Semantic Approaches for Knowledge Discovery and Retrieval in Biomedicine

Wilkowski, Bartlomiej

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Semantic Approaches for Knowledge Discovery and Retrieval in Biomedicine

Bartłomiej Wilkowski

Kongens Lyngby 2011
IMM-PHD-2011-261

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This thesis discusses potential applications of semantics to the recent literature-based informatics systems to facilitate knowledge discovery, hypothesis generation, and literature retrieval in the domain of biomedicine. The approaches presented herein make use of semantic information extracted from biomedical texts by natural language processing systems supported by biomedical ontologies. The thesis is divided into two main parts: first, a field of literature-based discovery is introduced, with a review of recent approaches of the field; second, literature retrieval in the domain of neuroimaging (neuroscience) is discussed with the emphasis put on the coordinate-based searching of related publications. My own contribution to the first part is a novel literature-based 'discovery browsing' methodology incorporating semantic predications, graph-based methods and path analysis for guiding researchers through the relevant literature on a user-specified biomedical phenomenon. Moreover, the additional analyses of the methodology show its potential application as a support for the recent probabilistic retrieval methods. In the second part of the thesis, I present the BredeQuery plugin which integrates a coordinate-based literature retrieval system with the common in neuroimaging statistical analysis environment. It is followed by the detailed description of a prototype of context-dependent neuroscientific literature retrieval methodology, which thanks to the employment of ontologies, allows the user to define context of interest for a search. The peer reviewed research articles, included in the appendices, discuss further the details of the presented methods, case studies, and provide other related information.

Resumé

Denne afhandling diskuterer mulige anvendelser af semantik i nyere litteraturbaserede informatik systemer til at søge ny viden, generere hypoteser, og finde litteratur i biomedicin. De metoder, der forelægges her gøre brug af semantisk oplysninger udtaget fra biomedicinske tekster ved hjælp af naturlig sprogbehandling systemer, der understøttes af biomedicinske ontologier. Afhandlingen er opdelt i to hoveddele: for det første et afsnit om litteraturbaseret opdagelse med en gennemgang af de seneste tiltag indenfor området; for det andet, indsamling af litteratur inden for neurobilleddannelse og neurovidenskab analyseret med vægt på koordinatbaseret søgning af relevante publikationer. Mit eget bidrag til den første del er en ny litteraturbaseret 'opdagelse browsing' metodologi baseret på semantisk prædikationer og grafteori der kan facilitere forskerne til at finde relevante litteratur om et specificeret biomedicinsk fænomen. Yderligere analyser af denne metode, indikerer ligeledes en potentiel anvendelse i forbindelse med probabilistiske metoder indenfor litteratursøgning. I den anden del af afhandlingen introduceres BredeQuery plugin, som integrerer et koordinatbaseret litteratursøgesystem i et neurobilleddannelse statistisk analysemiljø. Det efterfølges af en detaljeret beskrivelse af en prototype for kontekstafhængig neurovidenskabelig litteratursøgning, som ved hjælp af ontologier, giver brugeren mulighed for at definere rammerne af interesse for en søgning. Afhandlingens forskningsartikler, inkluderet i appendiks, diskuterer de nærmere detaljer om udformningen af de præsenterede metoder, casestudier, og andre praktiske detaljer.

Preface

This thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with different aspects of semantic analysis of the textual data with application to the biomedical domain. The main focus is on employing biomedical ontologies, semantic predications and natural language processing for retrieval of publications and literature-based knowledge discovery.

The thesis consists of a summary report and a collection of eight research papers written during the period 2008–2011, and elsewhere published.

Lyngby, June 2011

Bartłomiej Wilkowski

Papers included in the thesis

[A] Bartłomiej Wilkowski, Lars Kai Hansen. Context-dependent literature search: a support for functional imaging coordinate-based databases. *BMC Neuroscience*, 2011, (pp. 20). Submitted.

[B] Bartłomiej Wilkowski, Marcin Szewczyk, Peter Mondrup Rasmussen, Lars Kai Hansen, Finn Aarup Nielsen. BredeQuery: Coordinate-Based Meta-analytic Search of Neuroscientific Literature from the SPM Environment. *Part of: Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science: International Joint Conference, BIOSTEC 2009, Porto, Portugal, January 14-17, 2009, Revised Selected Papers*, 314–324, Springer-Verlag New York, 2010. Published.

[C] Bartłomiej Wilkowski, Marcelo Fiszman, Christopher M. Miller, Dimitar Hristovski, Sivaram Arabandi, Graciela Rosemblat, Thomas C. Rindflesch. Graph-Based Methods for Discovery Browsing with Semantic Predications. *American Medical Informatics Association Annual Symposium*, 1514–1523, Washington D.C., 2011. Published.

[D] Antonio Jimeno-Yepes, James G. Mork, Bartłomiej Wilkowski, Dina Demner Fushman, Alan R. Aronson. MEDLINE MeSH indexing: lessons learned from machine learning and future directions. *ACM SIGHIT International Health Informatics Symposium*, 2012, (pp. 5). Accepted.

[E] Antonio Jimeno-Yepes, Bartłomiej Wilkowski, James G. Mork, Elizabeth Van Lenten, Dina Demner Fushman, Alan R. Aronson. A bottom-up approach to MEDLINE indexing recommendations. *American Medical Informatics Association Annual Symposium*, 1583–1592, Washington D.C., 2011. Published.

[F] Bartłomiej Wilkowski. Neuroscientific literature search based on the location coordinates in brain - BredeQuery plugin for SPM environment. *Presented at: 2nd INCF Congress of Neuroinformatics, Front. Neur. Conference Abstract: Neuroinformatics 2009*, Pilsen, 2009. Published.

[G] Bartłomiej Wilkowski. Knowledge Discovery in Neuroinformatics. *Presented at: Medical Informatics in a United and Healthy Europe*, 150:589, Sarajevo 2009. Published.

[H] Bartłomiej Wilkowski, Marcin Szewczyk, Lars Kai Hansen. Bridging the gap between coordinate- and keyword- based search of neuroscientific databases by UMLS-assisted semantic keyword extraction. *Presented at: Human Brain Mapping*, NeuroImage, 47(S165):(pp. 3), San Francisco 2009. Published.

Acknowledgements

First, I want to thank my supervisor Lars Kai Hansen. Your constant support and inspirational force allowed me to develop my skills and broaden my knowledge significantly as your student during the last three years.

I want to thank the people from the Cognitive Systems section from DTU Informatics department, in particular to Carsten Stahlhut, Finn Årup Nielsen, Michael Kai Petersen, Jakob Eg Larsen, Marcin Szewczyk and Peter Mondrup Rasmussen. My special thanks to Ulla Nørhave for constant help given during my stay at DTU.

I want to express here my gratitude to Tom Rindflesch from the Lister Hill National Center for Biomedical Communications, NIH, USA for many fruitful discussions and very useful insights into my research work. Working with you and all the people from the Semantic Knowledge Representation project during my 5-month research visit was a great pleasure and unforgettable experience. I wish to thank also Lan Aronson, Francois Lang, Antonio Jimeno-Yepes, Marcelo Fiszman, Chris Miller, Graciela Rosemblat, Dongwook Shin, Han Zhang, Jim Mork, Sivaram Arabandi, Dina Demner Fushman.

I gratefully acknowledge the support of my sponsors: the DTU Informatics Graduate School ITMAN, the Center for Integrated Molecular Brain Imaging, the Otto Mønsteds Foundation, the Kaj og Hermilla Ostenfelds Foundation, and the Ingeniør Alexandre Haynman og hustru Nina Haynmans Foundation.

To my lovely Zosia, my beloved wife Dagmara, and my parents for their patience and permanent support. Without You I would not be where I am now.

Contents

Chapter 1

Introduction

Since the technological breakthrough in electronics, computers began to be an inseparable and indispensable elements of our daily life. More and more powerful computational machines allowed faster and more efficient research and development in many life sciences including medicine. Biomedical experiments started to consume much less time which allowed the implementation of various new analysis techniques and methods. The natural outcome of such a technological boom in biomedicine is the dramatically increasing production of scientific data.

Turning large amounts of scientific results and data into biomedical knowledge using the traditional "manual" meta analyses of results reported in journals and technical reports would be very time consuming, thus the resultant expansion of the medical databases has created a significant potential for the design of new data modeling and information retrieval tools and services that enable faster data processing, analysis, integration and dissemination among a highly interdisciplinary community of researchers (see Appendix G).

The field of information retrieval (IR) and various probabilistic and statistical approaches in machine learning have contributed to the problem of literature search and expanded significantly usage of literature databases. This resulted in creation of general publication search services like Google Scholar and also more specific, biomedical databases with search capabilities like the PubMed database.

Another possible way of dealing with textual data is by applying semantics, a science of meaning in language. Instead of treating text as a set of subsequent words and phrases incomprehensible to machines, in semantics, meaning of any text is inferred by computer through the provided structured domain knowledge sources called ontologies. 'Understanding' the meaning of documents by a computer system allows to design flexible retrieval systems which detect intentions of users and search in an appropriate context.

The pivotal aim of this thesis is to present semantic approaches for literature-based knowledge integration and manipulation, and discuss some of their possible applications in the biomedical domain, which may bring enhancement or support to the algorithms and methods recently employed for these purposes.

Semantic natural language processing

The approaches presented in this thesis make use of semantic information extracted from biomedical texts by natural language processing (NLP) systems supported by biomedical ontologies. A semantic predication, also known as a triple, is the smallest piece of information extracted by such NLP systems. It consists of two concepts (subject and object) related with each other via a carefully defined type of relationship, called predicate. The triple structure of such information may be easily stored in the suitable machine understandable data formats like RDF, OWL or even in well-constructed relational databases. Further processing of such data is much more effective than dealing with raw text.

Furthermore, semantic predications are also employed to provide visual summarizations of the knowledge encapsulated in a piece of text. A predication is visualized as two nodes connected with the arrow. The arrow represents a relationship (predicate) and the nodes represent two concepts involved in a relation. Figure 1.1 shows the example of how a sample biomedical text (title and abstract of a publication) may be visually represented using semantic predications. Such a clear visual summary allows very quickly to understand and identify paper's main ideas and topics.

The initial sections of Chapter 2 provide an insight into the process of semantic interpretation and predication extraction, as well as a review of two existing semantic NLP systems, BioMedLee (Lussier et al., 2006) and SemRep (Rindflesch and Fiszman, 2003).

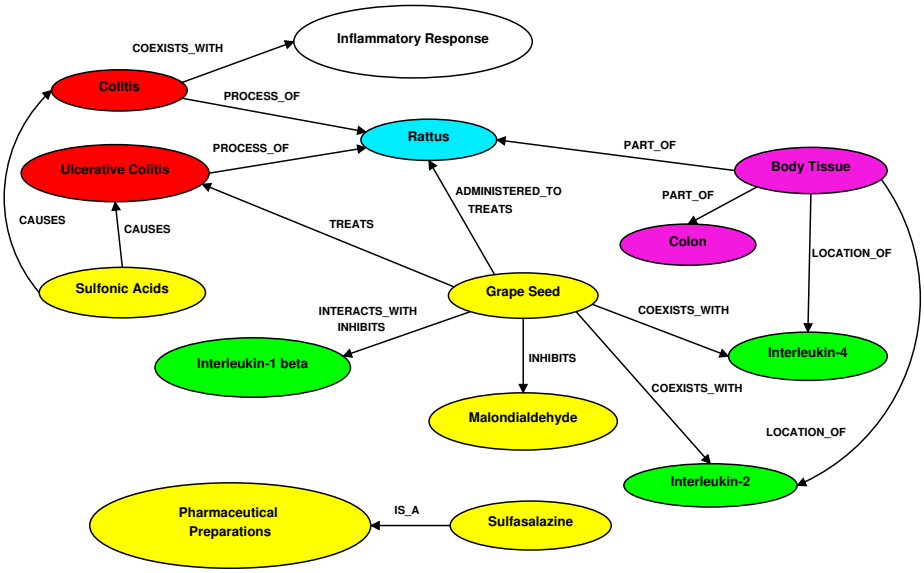


Figure 1.1: Visualization of semantic predications automatically extracted by the SemRep (Rindflesch and Fiszman, 2003) system from the title and abstract of the paper: "Therapeutic effect and mechanism of proanthocyanidins from grape seeds in rats with TNBS-induced ulcerative colitis." (Li et al., 2008). Different background colors of the biomedical concepts represent their semantic types: yellow (organic and inorganic chemicals), green (body substances), blue (animals), pink (body parts), red (diseases or syndromes) and white (body functions). Furthermore, the edges between the concepts represent type of the relationships, e.g.: substance interactions (INTERACTS_WITH, INHIBITS, COEXISTS_WITH), localization (PART_OF, LOCATION_OF), treatment of diseases (TREATS, CAUSES), etc.

Knowledge discovery in biomedicine

Krallinger and Valencia (2005) reviewed various literature retrieval and text mining tools for biomedical use and also described existing systems for so called 'knowledge discovery' in biomedical research. Pioneering methodology for finding unknown and implicit relations between biomedical concepts in huge repositories of biomedical literature was introduced by Swanson (1986a) and initiated a new field of research in information retrieval called literature-based discovery (LBD). Many LBD systems, including Swanson's approach, rely primarily on term co-occurrences without providing any semantic information about the type of the relation between concepts. Hristovski et al. (2006) expanded these approaches by employing semantic predications extracted from biomedical pub-

lications using ontology-based natural language processing systems.

The existing methods and approaches designed for LBD are discussed in Chapter 2. Moreover, I propose and discuss a 'discovery browsing' methodology, an extension for current LBD methodologies, employing graph theory and semantics. The underlying technology exploits the SemRep's (Rindflesch and Fiszman, 2003) semantic predications represented as a graph of interconnected nodes (predication arguments) and edges (predicates). The suggested paths in this graph represent chains of relationships. Consequently, this methodology guides the user through the research literature on a specified biomedical phenomenon (Appendix C).

Data integration in neuroscience

New, state-of-the-art scanning techniques caused a breakthrough in the development of neuroscience – a study of nervous system. Modern scanners allow to detect hemodynamic response, thus dynamically regulated blood flow in brain, relating directly to neuronal activity in brain, hence to define functional localization of specific human behavior (Ogawa et al., 1992). Functional localization is a dominant paradigm in current neuroimaging research and is usually reported as a set of coordinates, in reference to a specific brain atlas, representing a volume in brain.

There is an emerging need in neuroimaging for efficient integration of such structured datasets with ordinary literature databases. Despite that the existing coordinate-based literature databases are arguably richer for neuroimaging than conventional keyword-based retrieval services, the main challenge for this type of databases is to improve the labor intensive data entry process which has decreased the coverage and resulted in limited use. Consequently, to overcome this problem, new tools and solutions are needed to bridge the gap between coordinate- and keyword-based databases. The interconnection with more comprehensive bibliographical databases can extend the results pool of the coordinate-based services.

Figure 1.2 shows a sample activation data taken from a neuroimaging paper found through the web service of the Brede database (Nielsen, 2003). This database records published neuroimaging experiments that list stereotaxic coordinates in so-called MNI or Talairach space. Presently, close to 4000 coordinates from 186 papers with a total of 586 experiments are available. The Brede database provides similar visualizations to this from Figure 1.2 for all stored experiments.

x	y	z	Lobar anatomy
-4	63	15	Medial prefrontal cortex
-59	-5	-15	Left anteriolateral middle temporal gyrus
-48	10	-24	Left temporal pole
-22	-12	-11	Left hippocampus
-30	-36	-13	Left parahippocampal gyrus
-4	-53	25	Posterior cingulate gyrus
-48	-61	25	Left temporoparietal junction
51	-63	27	Right temporoparietal junction
0	-4	8	Thalamus

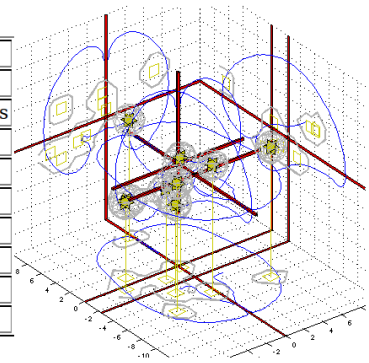


Figure 1.2: A sample set of activation coordinates for a neuroimaging experiment on memory retrieval: "Memory retrieval of temporal, nontemporal, person relevant and irrelevant memories by listening to statements and responding with key press versus listening sets of words and pressing a button depending on number of syllables in last word" (Maguire and Mummery, 1999) retrieved from the Brede database.

Chapter 3 deals with the current data integration needs in neuroimaging. Apart from a brief review of the coordinate-based databases and literature retrieval systems in neuroscience, I describe the BredeQuery plugin – an application which offers a direct link from the common neuroimaging environment Statistical Paremetric Mapping (SPM) to the Brede database (Appendix B). It provides a mechanism which allows the SPM user to find references to articles which relate to the similar brain activation areas. Finally, a methodology for a context-dependent literature retrieval using biomedical ontologies and natural language processing is discussed (Appendix A).

Contributions

Five full papers (Appendices A–E) and three short papers (Appendices F–H) included in this thesis refer to the biomedical domain, and are classified into three categories. Appendix C deals with a methodology for the field of literature-based discovery. Appendices A, B, F, G, H are oriented towards semantic approaches for integration and expansion of the coordinate-based services in neuroscience. Appendices E and D refer to machine learning approaches for improving MeSH indexing of MEDLINE publications and are not further discussed herein since they go beyond the main subject of this thesis.

Chapter 2

Literature-based discovery of biomedical knowledge

The main thrust of this chapter centres around the field of literature-based discovery (LBD). First, the notion of semantic predications is introduced in Section 2.1, since that approach provides the strongest basis for LBD. The review of current semantic predication extraction systems precedes Section 2.2, which discusses approaches for knowledge summarization using predications. Then, the background and principles of LBD is presented in Section 2.3, followed by a review of various LBD methodologies, including the approaches based on the semantic predications (Section 2.4). Finally, I present in Section 2.5 a novel methodology, based on semantic predications and designed using graph-based methods and path analysis, which extends the existing LBD approaches.

2.1 Semantic predications

The automatic extraction of ontological relations between various concepts from biomedical texts is a complicated process recently established by just few systems. One such natural language processing system, BioMedLee (Chen and Friedman, 2004; Lussier and Friedman, 2007), uses various biomedical ontologies like Unified Medical Language System (UMLS), Mouse Genome Informatics

(MGI) or Mammalian Ontology to extract many different types of phenotypic relationships. It is incorporated into the PhenoGO system which automatically augments annotations in Gene Ontology annotations with additional context (Lussier et al., 2006). BioMedLee is dedicated to the gene-related domains, thus it may bring potential enhancement to LBD by integration with other LBD systems. Hristovski et al. (2006) integrated BioMedLee with another semantic relation extraction system – SemRep (further discussed in Section 2.1.1) and BITOLA LBD system for evaluation of drug-disease discovery pattern (Section 2.4).

2.1.1 Extraction with SemRep system

Another system, which provides the functionality of semantic information extraction from the biomedical domain is SemRep system (Rindflesch and Fiszman, 2003). Figure 2.1 presents the pipeline of subsequent steps followed during the process of extraction of semantic predications from biomedical texts.

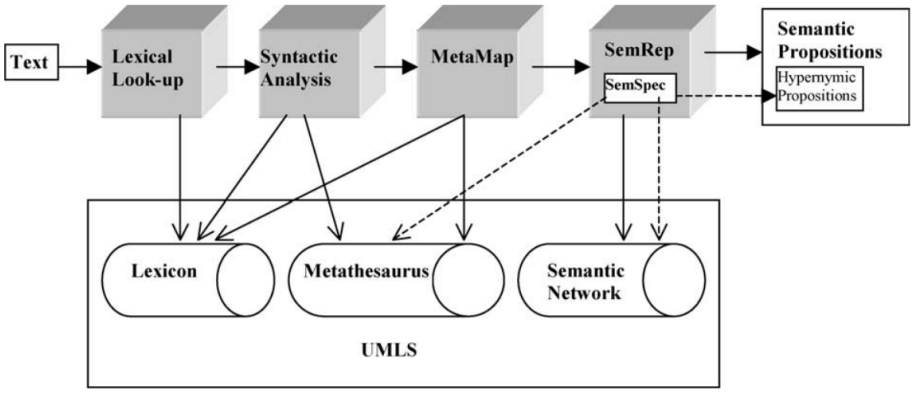


Figure 2.1: The architecture of the natural language processing system for semantic predication extraction (Rindflesch and Fiszman, 2003). The whole system relies on the resources of Unified Medical Language System (UMLS). Given a biomedical text, lexical and syntactic analysis is followed by the concept mapping step by Metamap program (Aronson and Lang, 2010). The last phase of this pipeline is performed by the SemRep system, which assigns semantic relationships from UMLS's Semantic Network to various pairs of UMLS's Metathesaurus concepts detected by Metamap.

The SemRep system is developed by Thomas C. Rindflesch group at the Lister Hill National Center for Biomedical Communications at the U.S. National

Library of Medicine. SemRep extracts semantic predications (triples), see Figure 2.2, from biomedical texts. The output is stored in the relational database called SemMed to allow their further processing. The recent version of SemMed database consists of almost 27 million semantic predications extracted from almost 8 million MEDLINE papers published after January 1, 1999.

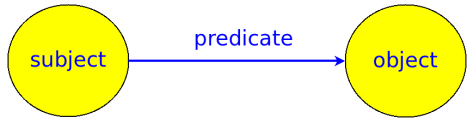


Figure 2.2: A semantic predication model. It is a statement of a relationship between two concepts represented in a from of a triple. In Semantic Web, it is an underlying (triple) structure of any expression representing information and stored in Resource Description Framework (RDF). Two concepts of a semantic predication (subject and object) are in a relationship (predicate) represented by a directed arc.

As it is shown on Figure 2.1, the biomedical text is analyzed by the Metamap program. Metamap (Aronson and Lang, 2010) is a software based on natural language processing and computational linguistics. It maps simple noun phrases to UMLS Metathesaurus concepts. The natural language processing is performed using MedPost tagger, which achieves over 97% accuracy on MEDLINE citations (Smith et al., 2004a). Metamap output includes metadata of text phrases such as part of speech, mapped concepts identifiers and labels, preferred concepts, semantic types and knowledge sources for mapped concepts. SemRep gets the concepts together with assigned semantic types as the input. Let's consider the following sample sentence (title of a paper):

> "Perospirone augmentation of paroxetine in treatment of refractory obsessive-compulsive disorder with depression." (Otsuka et al., 2007).

Metamap detects UMLS concepts and their semantic types. Some of them are:

- perospirone [Organic Chemical,Pharmacologic Substance]
- Augmentation (Augmentation procedure) [Therapeutic or Preventive Procedure]
- Paroxetine [Organic Chemical,Pharmacologic Substance]
- Refractory (Unresponsive to Treatment) [Functional Concept]
- Obsessive-Compulsive Disorder [Mental or Behavioral Dysfunction]

Later, using Metamap's output and performing additional lexical analyses and rule matching, SemRep returns three semantic predications (subject – predicate – object):

1. Augmentation procedure – USES – perospirone
2. perospirone – STIMULATES – Paroxetine
3. Augmentation procedure – TREATS – Obsessive-Compulsive Disorder

There are 48 different predicates (INHIBITS, TREATS, etc.) used by SemRep accompanied by 43 negation predicates (NEG_INHIBITS, NEG_TREATS, etc.). SemRep maps syntactic indicators (verbs, prepositions, nominalizations, etc.) to predicates in UMLS Semantic Network. For example, preposition *of* maps to predicate USES in predication 1, nouns *augmentation* and *treatment* map to STIMULATES and TREATS in predications 2 and 3 respectively. Moreover, each predicate has the "permissible argument configurations" which define sets of semantic types for UMLS Metathesaurus concepts which are allowed as subject and object in a predication triple (Rindflesch and Fiszman, 2003). The example of few permissible argument configurations for various predications are presented in Table 2.1.

Subject	Predicate	Object
Therapeutic or Preventive Procedure	USES	Pharmacologic Substance
Pharmacologic Substance	STIMULATES	Pharmacologic Substance
Organic Chemical	STIMULATES	Gene or Genome
Pharmacologic Substance	TREATS	Disease or Syndrome

Table 2.1: Sample permissible argument configurations for predicates USES, STIMULATES and TREATS in SemRep system. If there is a mapping between a syntactic indicator and a predicate in a natural language fragment, and then subject and object fulfill at least one permissible argument configuration for this predicate, SemRep defines it as a valid semantic predication.

2.2 Knowledge summarization with predications

To complete the view on semantic predications before moving to the field of literature-based discovery, it is worth to mention the Semantic MEDLINE[1] (Kilicoglu et al., 2008), a web-service that summarizes and visualizes semantic predications extracted by SemRep system for user's PubMed searches. It nicely

[1] http://skr3.nlm.nih.gov/SemMedDemo/

demonstrates how semantic relations may improve selection and retrieval of relevant papers.

While Figure 1.1 in Chapter 1 visualized only one given biomedical paper with semantic predications, Semantic MEDLINE is able to display similar visualizations for a PubMed output (set of papers), see Figure 2.3. These views are summarized, thus they represent a flavor of the given topic, showing only its most important aspects and disregarding uncommon, infrequently occurring information.

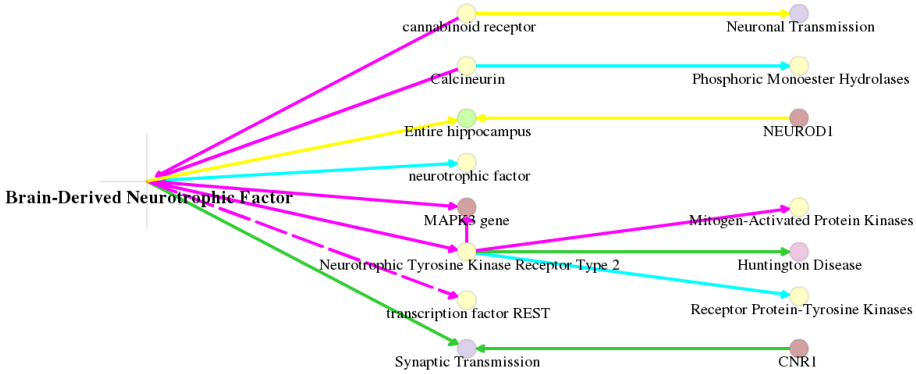


Figure 2.3: Summarization graph in Semantic MEDLINE – Substance Interactions perspective on Huntington Disease for 500 most recent papers in MEDLINE (topic retrieved with a query: Huntington's disease[mh]). Colors of arcs represent predicates: DISRUPTS (yellow), AFFECTS (green), INTERACTS_WITH (purple), ISA (cyan). Each arc represents a predication extracted from title and abstract of one or more MEDLINE papers.

Semantic MEDLINE's user selects a topic (defined by a PubMed query) and a summarization perspective. A summarization perspective consists of a set of allowed predications (ontological predicates together with the permitted classes of concepts for subject and object). There are recently four different summarization perspectives available:

- Treatment of Disease (Fiszman et al., 2004):
  - {Disorders} ISA {Disorders}
  - {Etiological process} CAUSES {Disorders}
  - {Treatment} TREATS {Disorders}
  - {Body location} LOCATION_OF {Disorders}
  - {Disorders} OCCURS_IN {Disorders}

  - {Disorders} CO-OCCURS_WITH {Disorders}

- Substance Interactions (Fiszman et al., 2006):
  - {Drugs} AFFECTS {Disorders}
  - {Drugs} AFFECTS {Physiology}
  - {Drugs} CAUSES {Disorders}
  - {Drugs} COMPLICATES {Disorders}
  - {Drugs} COMPLICATES {Physiology}
  - {Drugs} DISRUPTS {Anatomy}
  - {Drugs} DISRUPTS {Physiology}
  - {Drugs} INTERACTS_WITH {Chemicals}
  - {Drugs} ISA {Chemicals}
  - {Drugs} PREVENTS {Disorders}
  - {Drugs} TREATS {Disorders}
- Diagnosis (Sneiderman et al., 2006):
  - {Etiological process} CAUSES {Disorders}
  - {Diagnostic process} DIAGNOSES {Disorders}
  - {Anatomy} LOCATION_OF {Disorders}
  - {Disorders} CO-EXISTS WITH {Disorders}
  - {Disorders} PROCESS_OF {Living Being Human}
  - {Disorders} ISA {Disorders}
- Pharmacogenomics (Ahlers et al., 2006):
  - {Substance} ASSOCIATED_WITH {Pathology}
  - {Substance} PREDISPOSES {Pathology}
  - {Substance} CAUSES {Pathology}
  - {Substance} INTERACTS_WITH {Substance}
  - {Substance} INHIBITS {Substance}
  - {Substance} STIMULATES {Substance}
  - {Substance} AFFECTS {Anatomy OR Process}
  - {Substance} DISRUPTS {Anatomy OR Process}
  - {Substance} AUGMENTS {Anatomy OR Process}
  - {Substance} ADMINISTERED_TO {Living Being}
  - {Process} MANIFESTATION_OF {Process}
  - {Substance} TREATS {Living Being OR Pathology}
  - {Anatomy OR Living Being} LOCATION_OF {Substance}

- {Anatomy} PART_OF {Anatomy OR Living Being}
- {Process} PROCESS_OF {Living Being}
- {Substance} CO-EXISTS_WITH {Substance}
- {Process} CO-EXISTS_WITH {Process}

The visualizations in Semantic MEDLINE, as presented in Figure 2.3, are directed graphs (networks) representing biomedical concepts and their semantic connections. In some cases of widely studied diseases, the size of the graph, thus number of predications, constructed by Semantic MEDLINE is too large and dense to be easily analyzed by users. Recent initiative by Zhang et al. (2011) employs graph-based method for enhancing Semantic MEDLINE summarization. The degree centrality measure is used as a threshold for selecting concepts that are related to the topic and highly interconnected in the semantic predications graph.

Following the two subsequent sections which introduce the field of literature-based discovery (Section 2.3) and describe current applications to that field (Section 2.4), we propose in Section 2.5 a novel LBD methodology incorporating semantic predications and graph theory in order to guide researchers through the relevant literature on a user-specified biomedical phenomenon.

2.3 Principles of literature-based discovery

The result of an efficient literature-based discovery approach is a new, or poorly studied, knowledge which may lead to a discovery. In biomedicine, a discovery of new relations between various body substances and their responses to drugs is a crucial step towards better understanding of sophisticated body functions and mechanisms, what potentially may lead to inventions of more effective cures for various diseases. Various literature-based discovery approaches in biomedicine analyze huge corpora of biomedical and life sciences publications in order to reveal implicit relations.

Swanson (1986b) introduced the idea of two complementary sets of articles, which when studied together can reveal useful scientific information impossible to be asserted when studying them separately. Such two sets of articles are assumed to be isolated from each other by not being cited together and by each other. In other words, the two complementary sets refer to different domains.

Such inter-domain and synergistic paradigm was presented by Swanson (1986a) on the example of dietary fish oils and Raynaud's disease. The first set of papers

analyzed by Swanson discussed that dietary fish oils lead to certain blood and vascular changes. The second set of papers mentioned that the same blood and vascular changes are beneficial for patients with Raynaud's disease. It was then inferred and hypothesized that dietary fish oils may help in treating Raynaud's disease, what was never before discussed in any research paper. After few years, first clinical trials validated the above mentioned discovery.

Swanson formalized and defined two types of literature-based discovery: open discovery and closed discovery. In open discovery paradigm, the basic underlying principle is that relations A – B and B – C may be known, yet relation A – C has gone unnoticed. For instance, knowing the biomedical problem A, e.g. Raynaud's disease in Swanson's example, we try to find different related concepts B, like body processes, interacting with A. Blood viscosity, platelet aggregation and vascular reactivity are few of many examples of B concepts. Then, a set of concepts C interacting with the selected group of Bs is to be identified. In Swanson's example it led to a dietary fish oils as C. Open discovery paradigm is also called as a hypothesis generation process since an implicit relation of A (Raynaud's disease) interacting with C (dietary fish oils) was generated through the described step-by-step process involving concepts B (e.g. blood viscosity).

Figure 2.4: Hypothesis generation (open discovery) and hypothesis testing (closed discovery) paradigms. Open discovery assumes that starting with a selected biomedical issue A, its potential relationship with concepts C may be found indirectly through the detailed analysis of middle concepts B. Closed discovery assumes that the relation A – C is known and then such a hypothesis is tried to be proven by finding appropriate intermediate links B.

In closed discovery paradigm, the basic underlying principle is that relation A – C may be known, yet relations A – B and B – C have gone unnoticed. Closed discovery paradigm is often described as hypothesis testing process since the hypothesis (A – C) is known before the discovery process begins. To demonstrate it again on Swanson's example, a researcher assumes a relationship between Raynaud's disease (A) and dietary fish oils (C) as known (hypothesis). Then, a set of concepts B is to be found to conclude the discovery process. Here it is done by finding blood viscosity or vascular reactivity concepts (B) as links between A and C. Figure 2.4 visualizes the open and closed discovery paradigms.

Swanson continued the development of his pioneer LBD approaches (Swanson and Smalheiser, 1996) what resulted in the release of Arrowsmith (Swanson and Smalheiser, 1997) computer system and online service[2] designed to assist open discoveries in biomedicine. Moreover, Swanson's foundations stimulated other biomedical researchers to develop other LBD systems (Gordon and Lindsay, 1996; Weeber et al., 2000; Hristovski et al., 2001; Weeber et al., 2001; Srinivasan and Libbus, 2004; Fuller et al., 2004; Hristovski et al., 2005). The functionality of all of these systems and approaches, including Swanson's, employed, as a main mechanism, cooccurence of terms or concepts found in titles and abstracts of biomedical literature. Hristovski et al. (2006) concluded that "the use of co-occurrence has several drawbacks, since not all co occurrences underlie 'interesting' relations: (a) Users must read large numbers of Medline citations when reviewing candidate relations; (b) systems tend to produce large numbers of spurious relations; and, finally, (c) there is no explicit explanation of the discovered relation.". There were further attempts discussed by Cole and Bruza (2005), introducing "dimensional reduction", thus a decrease in the number of candidate relations for LBD approaches, using statistical and probabilistic methods like latent semantic indexing or singular value decomposition. Nevertheless, still none of these methods reflects semantic nature, or meaning, of relations between concepts.

To enhance LBD systems and provide semantic information about type of relations, the natural language processing systems for extraction of semantic predications from literature were developed. Section 2.1 introduced the idea of extraction of semantic information from biomedical texts. The next section reviews applications of semantic predications in literature-based discovery.

2.4 Application of semantic predications in literature based discovery

This section summarizes the applications of SemRep's semantic predications to literature-based discovery. A broader insight into the recent advances in LBD using natural language processing and semantics is given by Hristovski et al. (2008).

In parallel with addressing issues with co-occurrence as a primary mechanism in literature-based discovery, Hristovski et al. (2006) proposed a refinement for focusing on useful relations by employing semantic predications from SemRep. Based on Swanson's A – B – C approach, they introduced the notion of *discovery*

[2] http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

pattern which can be understood as a defined set of conditions to be fulfilled for a discovery. It is illustrated on the open discovery *Maybe_Treats* pattern, relying on the relation TREATS from SemRep system, which says (in part) that a therapeutic agent C maybe treats disease A if the level of an important measurement B is typically increased in patients with disease A and if C is able to reduce the level of B. To evaluate this discovery pattern, the Swanson's Raynaud's discovery was replicated using the BITOLA[3] (Hristovski et al., 2001, 2005) LBD system.

The *Maybe_Treats* pattern was later modified into *inhibit the upregulated* and *stimulate the downregulated* patterns assuming genes placed as B concepts in drug-disease A – B – C discoveries (Hristovski et al., 2010). Authors integrated the semantic predications database with DNA microarray results. Results of the analysis done on Parkinson's disease propose various hypotheses including substances potentially effective in its treatment.

Another discovery pattern, *May_Disrupt* (see Figure 2.5), concentrates on pharmacogenomics, thus relationship among drugs, genes and diseases (Ahlers et al., 2007). It tries to discover a substance A which may potentially prevent or treat the disease C by finding the cause of the disease C (substance B). Using this pattern, the authors explicate, through the closed discovery, the mechanism of the antipsychotic drugs therapy (A) on cancer (C). This analysis revealed five bioactive substances (B) reported in literature as both prone to inhibition by antipsychotic agents and involved in the etiology of cancer: brain-derived neurotrophic factor, CYP2D6 gene, glucocordicoid receptor, PRL gene, and TNF gene.

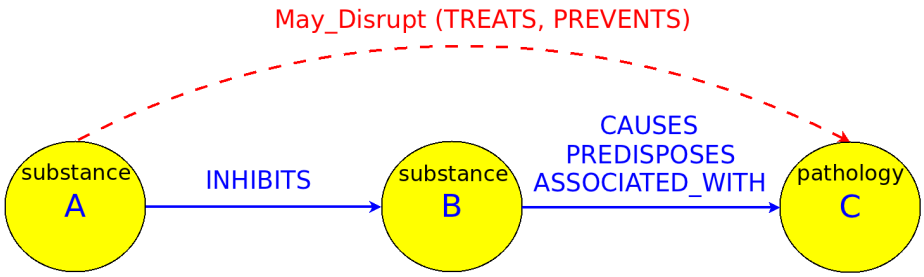


Figure 2.5: *May_Disrupt* discovery pattern.

Finally, Cohen et al. (2010b) presented the EpiphaNet LBD system, which derives all the knowledge from SemMed database of semantic predications. The system is able to predict term co-occurrence and as a result may simulate open and closed discoveries. EpiphaNet employs methods of distributional semantics

[3]http://www.mf.uni-lj.si/bitola/

and finds relations between concepts which do not co-occur directly using Reflective Random Indexing (RRI) (Cohen et al., 2010a). Moreover, it makes use of Predication-based Semantic Indexing (PSI) (Cohen et al., 2009), a distributional model of language for representing semantic predications in a compact vector space. EpiphaNet, through RRI and PSI, may be thought as a novel way for information retrieval and discovery providing complementary measure of semantic similarity between terms or concepts.

2.5 Discovery browsing methodology

The methodology described here goes beyond making discoveries to a principled way of navigating through selected aspects of some biomedical domain. The method is a type of "discovery browsing" that guides the user through the research literature on a specified phenomenon. One of the aims is to go beyond document retrieval in response to a query by revealing crucial relationships in the domain, which may evolve as the user exploits the method. Poorly understood relationships may be explored through novel points of view, and potentially interesting relationships need not to be known ahead of time. In a process of "cooperative reciprocity" the user iteratively focuses system output, thus controlling the large number of relationships often generated in literature-based discovery systems. The underlying technology exploits SemRep semantic predications represented as a graph of interconnected nodes (predication arguments) and edges (predicates). Later, the system suggests paths in this graph, which represent chains of relationships.

The following section introduces graph-based methods used in the described approach. Next, a brief description of the methodology followed by discussion of results from a biomedical study is given. The general motivation, methodology description and results from the depressive disorder study is to be found in Appendix C. Further analyses concerning this study is covered in Section 2.5.3.

2.5.1 Graph-based methods for discovery browsing

A graph is a representation of connections (edges) between objects (nodes). Graphs, also known as networks, are extensively studied in social network analysis and the Semantic Web. In our case where graphs are built of semantic predications, any edge represents a predicate, and any node is a concept. Graph theory is a set of functions and measures pertaining to graph properties. One such measure used in our methodology is degree centrality, which measures the

connectedness of nodes in a graph. A node with more connections (relationships) to other nodes has higher degree centrality. Freeman (1979) describes degree centrality as an indicator of the communication activity in a social network, which in our case may be considered as an indicator of the principal biomedical concepts in the domain for which the graph was constructed. The formula for degree centrality of node v in a graph with n nodes is

$$C_d(v) = \frac{deg(v)}{n-1}, \tag{2.1}$$

where $deg(v)$ represents a degree of node v thus a number of connections to other nodes.

We introduce also the measure of weighted degree centrality of a node. In our graphs, each edge has a number assigned representing a count of sentences from a corpus of MEDLINE papers in which a given predication was found. This count represents a weight of a predication. The formula for weighted degree centrality of node v in a graph with n nodes and weight w_i of ith edge connected to node v is

$$C_{wd}(v) = \sum_i w_i. \tag{2.2}$$

In graph theory, a path is a sequence of edges connecting any two nodes in the graph. Paths may be of any length. The shortest is of length 1:

$$A - B. \tag{2.3}$$

The longest is of length $N-1$, where N is the number of nodes in the graph

$$X_1 - X_2 - ... - X_N. \tag{2.4}$$

In Semantic Web research on ranking paths of semantic associations, Anyanwu et al. (2005) exploit the notion of "predictability." In their results longer paths more likely reveal rare and uncommon associations.

Dupont et al. (2006) discuss many walking approaches in a graph (edge passages), which may be also understood as extraction of paths from the graph. The definitions of maximal length of the edge passage (k-walk) and nodes of interest are based on this work. The nodes of interest are the start and end points

of a walk in a graph. For them, length of the walk is the number of intermediate nodes visited during a walk between nodes of interest. We measure path length by the number of edges between the start and end nodes.

2.5.2 Methodology description

This section describes subsequent steps of the presented "discovery browsing" methodology (Figure 2.6).

Seed definition – input

First, the user specifies a seed concept (or concepts), which describe the domain of interest to be analyzed. All the predications from the whole SemRep predication database are extracted which involve the seed. In addition, a list of predicates of interest (predicate pool) may be specified to limit the extracted predications to a specific need.

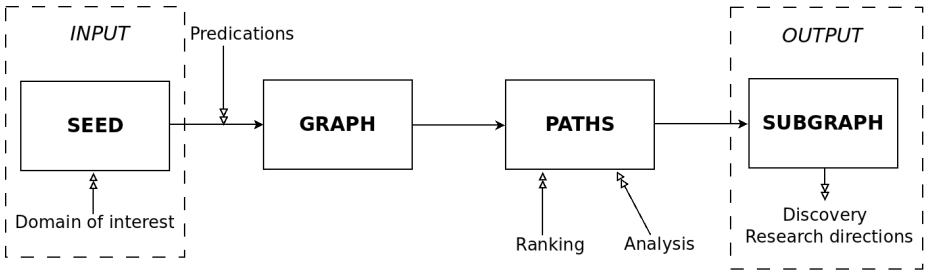


Figure 2.6: Architecture of discovery browsing methodology.

Graph creation

The predications are loaded into initial graph. Concepts in the graph are ranked by degree centrality, and those highest on the list, are used to extract additional predications to be added to the growing graph. This process is iterative and is continued until a graph of sufficient size to produce "interesting" results has been generated.

Path analysis

The process of extraction of various paths from the final graph is performed. The user defines anchors and length of the paths to be extracted. Finally, the paths are ranked, based also on degree centrality or weighted degree centrality measure.

Subgraph – output

The user selects a set of paths for further analysis from which a small subgraph is built. After the detailed analysis of relatively small set of papers assigned to the predications in the subgraph, a new hypothesis or research direction in the domain of interest is to be defined.

The assumption of this methodology is to allow the user to control the execution of the subsequent steps and change earlier defined parameters of the analysis (seed concepts, predicate pool, type of concept or path ranking) at any time if desired.

2.5.3 Depressive disorder study

The discovery browsing methodology was employed in a practical case, where unknown or poorly studied substance interactions in depressive disorder were to be found. The outcome of this study was discussed and evaluated by three biomedical experts. Moreover, the accuracy of SemRep system was evaluated by the expert in linguistics. The results are presented in the Appendix C. Here I summarize these results and present additional study-related demonstrations.

The domain of interest in this analysis was depressive disorder. Serotonin was selected as seed, since it is known to be a prominent neurotransmitter in this disorder. After the execution of the methodology, a subgraph indicating interaction of circadian rhythms, melatonin, proinflammatory cytokines and norepinephrine was striking.

The detailed manual analysis (carried out by field experts) of the retrieved papers revealed three major components of our results: 1) inflammation and depression, 2) circadian phenomena and depression, 3) noradrenergic aspects of depression. The additional PubMed queries showed that even varying amounts of research have been devoted to each of these components, little has considered all three together. Our results do not constitute a discovery in the sense

of something previously not noticed by anyone. However, in several respects they contribute to various aspects of depression that are currently incompletely understood and have not been extensively studied. Insight into the interrelationships among all these components may materially contribute to unraveling the underlying pathophysiology of depression, thus underpinning more effective treatment (and prevention).

The outcome of this analysis is a small subgraph of relations between five biomedical concepts: CLOCK gene, melatonin, interleukin-6, interleukin-1 beta, and norepinephrine (Figure 2.7). There are three major relationship segments in the subgraph:

- CLOCK gene – melatonin
- melatonin – proinflammatory cytokines
- proinflammatory cytokines – norepinephrine

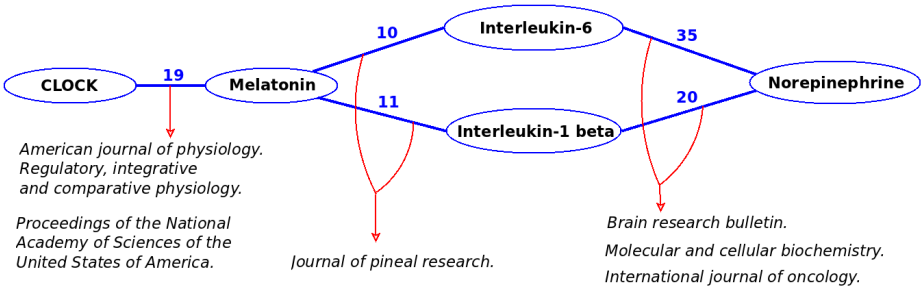


Figure 2.7: Depression disorder study output.

Figure 2.7 associates also each subgraph's segment with journal names in which the major amount of segment's papers were published. It creates a domain-wise classification of the subgraph's segments. While the majority of predications consisting on the *CLOCK gene – melatonin* segment belong to the physiology domain (American journal of physiology), the *melatonin – proinflammatory cytokines* segment consists of majority of papers from the pineal research (Journal of pineal research). Finally, the predications in the last segment, *proinflammatory cytokines – norepinephrine*, were published in a diverse set of biomedical domains starting from the brain research, via molecular and cellular biochemistry to oncology (International journal of oncology).

To conclude, the methodology provides the user (biomedical researcher) the ability to navigate through topic-related studies published in various biomedical

domains, some of which may go beyond the normal scope of interest of the user, which consequently may lead to new hypotheses and discovery of new knowledge.

2.5.4 Potential enhancements to literature retrieval

Most current document retrieval systems rely on probabilistic and statistical methods with a reasonable level of accuracy. Such systems take a query with keywords of interest as input, and based on their distribution among a huge number of papers, retrieve and rank those which best fit the given query. Unfortunately, sometimes, even a well constructed query, results in a large number of retrieved papers, not all of which are relevant. This is caused by the fact that the user has not been provided with a straightforward way of defining what aspect of the retrieved papers should match the query. Similarly unsatisfactory results are often produced by document neighboring systems, since the user is not able to define particular characteristics of the query document that should be included in the output documents.

The analysis below proposes an enhancement of current retrieval systems by combining them with advanced filtering using semantic predications, which allows the user to to specify in detail the topic, fact or relation of interest. This may be thought of as narrowing the context of search. The analysis is based on the papers retrieved by the discovery browsing methodology described in Section 2.5.3.

In the final subgraph retrieved in the discovery browsing methodology for the depressive disorder case, there are five different relations between five different biomedical concepts. In this analysis we concentrate on only one of these: the interaction between the two substances, CLOCK gene and melatonin. Based on a very large corpus of MEDLINE citations (titles and abstracts), the subgraph shows that this relation was extracted from 19 sentences found in 16 MEDLINE citations.

The 16 citations having predications asserting an interaction between CLOCK gene and melatonin constitute a coherent set of documents about this relationship. Given any one of these it is easy to find the other papers about this interaction by matching the relevant predications. In this experiment we compared that processing to the use of the PubMed Related citations feature, a type of statistically based document neighboring algorithm.

We first retrieved through PubMed Related the related citations for each of the 16 query papers (those having CLOCK gene-melatonin predications). We then examined these to see how many of the related citations discussed the relevant

	Found in PubmedRelated output?																
Input paper (seed)	16488540	12111547	16838899	16840546	20404168	19087173	11490096	16441555	12111535	10971628	18662218	12521925	14972686	12374857	11959680	15913571	Coverage (%)
16488540	•	•		•	•	•			•								37.5
12111547	•	•	•		•	•			•			•					43.8
16838899		•	•			•											18.8
16840546	•			•		•					•						25.0
20404168	•	•			•	•											25.0
19087173	•	•	•	•	•	•		•	•								50.0
11490096							•	•									12.5
16441555						•	•	•	•	•	•	•	•				50.0
12111535	•	•				•		•	•								31.3
10971628								•		•				•			18.8
18662218				•				•			•						18.8
12521925		•						•				•					18.8
14972686								•					•	•			18.8
12374857										•			•	•			18.8
11959680															•	•	12.5
15913571															•	•	12.5

Table 2.2: The results of the PubMed Related output analysis for all the 16 papers from the final subgraph of the depression disorder study in which a relation between CLOCK gene and melatonin was found. Each of the 16 papers (left column) was input to PubMed Related and the (output) list of related publications was retrieved. Then, it was checked which of the 16 papers were found in the PubMed Related output. The found papers are marked with a bullet symbol (•).

substance interaction asserted in the query document. Table 2.2 summarizes how many of the 16 query papers are found in the list of related publications returned by PubMed Related for a given seed paper. It shows that PubMed Related citations algorithm was able to find on average only a quarter of the 16 papers (in the two best cases, half) determined by the subgraph as being relevant to the CLOCK-Melatonin relationship. We have carried out a more detailed analysis of the two best cases, related citations returned for seed citations PMID 19087173 (Table 2.3) and PMID 16441555 (Table 2.4).

In the analyses we studied the top 10 papers (excluding seed paper) returned by PubMed Related. In case of PMID: 19087173 (Table 2.3), we have noticed 2 false negatives (FN), which means that the two papers (positions #3 and #11) mention in their title and/or abstract a relation between CLOCK gene and melatonin, but were not extracted by SemRep system. There are also 5 true negatives (TN), which means that even though these papers are generally related to the seed paper, they do not mention any relation between CLOCK gene and melatonin (correct performance of SemRep). The remaining 3 papers were published before 1999, and thus they were not processed by SemRep. Nonetheless, we manually analyzed the abstracts of these papers and all of them are TNs. In case of PMID: 16441555 (Table 2.4), there are 8 citations published before 1999 thus not processed by SemRep. Processing these by hand reveals that they are all TN, in that none discusses a relation between CLOCK gene and melatonin. There is also one FN on position #7.

It has to be clarified here that TN (true negative) does not necessarily mean that a paper is not relevant, but that it does not specifically discuss the CLOCK gene and melatonin interaction. On the other hand, true positive (TP) means that such interaction is discussed in a paper and that it was correctly found by SemRep. Finally, FN (false negative) indicate a case when a paper discusses the CLOCK gene and melatonin interaction, but it was not detected by SemRep (incorrect performance).

In the above tables one can notice that the subgraph papers, which were found by PubMed Related are ranked relatively low on the papers list. This analysis shows the potential enhancement of current probabilistic retrieval and ranking methods – PubMed Related uses one of such algorithms (Lin and Wilbur, 2007) – with semantic approaches. Current PubMed Related algorithm is undoubtedly very efficient and retrieves papers similar to the seed paper. The issue is that it does not take into account the way in which one paper relates to the other. PubMed Related, as well as other similar systems, rank and retrieve papers based on the general co-ocurrence of terms.

As an example specifying a relationship, in Table 2.4 we have looked for papers related to the paper (16441555): “Trans-pineal microdialysis in the Djungarian

#	PMID	Year	Status	Comment
1	19087173	2008	Seed	"In mammals, the rhythmic synthesis of melatonin by the pineal gland is tightly controlled by the master Clock located in the suprachiasmatic nuclei (SCN)."
2	12542658	2003	TN	The abstract or title of this paper does not refer to CLOCK gene.
3	16709646	2006	FN	"One of these genes, Clock, has been considered essential for the generation of cellular rhythmicity centrally and in the periphery; however, melatonin-proficient Clock(Delta19) + MEL mutant mice retain melatonin rhythmicity, suggesting that their central rhythmicity is intact."
4	15703405	2005	TN	Neither the abstract nor the title of this paper indicate relationship between CLOCK gene and melatonin.
5	15193530	2004	TN	Even the paper mentions relationships betwen melatonin and suprachiasmatic nuclei, PER2 gene and Arg-vasopressin, it does not indicate the relationship of interest: CLOCK gene – melatonin.
6	20876815	2010	TN	Neither the title nor the abstract of this paper refer to melatonin.
7	7632982	1995	TN	Neither the abstract nor the title of this paper indicate relationship between CLOCK gene and melatonin.
8	9886827	1999	TN	Neither the abstract nor the title of this paper indicate relationship between CLOCK gene and melatonin.
9	8274765	1993	TN	Even the paper indicates relationship between melatonin and circadian rhythms, CLOCK gene is not mentioned.
10	16687310	2006	TN	Neither the title nor the abstract of this paper mention CLOCK gene.
11	15590161	2005	FN	"The daily rhythm of melatonin synthesis in the rat pineal gland is controlled by the central biological clock, located in the suprachiasmatic nucleus (SCN) . . . "
. . .				
177	16840546	2006	TP	"Maternal melatonin effects on clock gene expression in a nonhuman primate fetus."
194	16838899	2006	TP	"The pineal hormone melatonin is a specific and sensitive marker of the circadian clock activity, and its secretion is tightly coupled to the output of the biological clock and The circadian phase."
203	12111547	2002	TP	"This interaction, which grants a temporally precise regulation of gene expression, may reflect the central role of melatonin, i.e. in synchronising peripheral clock cells that require unique phasing of output signals with the master clock in the brain."
230	16488540	2006	TP	"Pinealectomy has no effect on clock gene rhythms in the suprachiasmatic nucleus (SCN), the master circadian clock, as well as in the eyes and heart, indicating that the effect of melatonin on clock gene rhythms is tissue specific."
293	20404168	2010	TP	"These results demonstrate that melatonin suppresses the Clock/+ mutant phenotype and interacts with Clock to affect the mammalian circadian system."
337	12111535	2002	TP	"Finally, the bases for acute and clock regulation of the key enzyme in melatonin synthesis, arylalkylamine N-acetyltransferase (AA-NAT), are described."
394	16441555	2006	TP	"The rhythmic secretion of melatonin by the pineal gland is under control of the circadian clock, conveying the photoperiodic message to the organism."
. . .				
1142				

Table 2.3: Analysis of the top 10 (excluding seed) related citations for a seed paper (in yellow): PMID 19087173, retrieved by Pubmed Related. Papers highlighted in orange are those from the set of 16 papers analyzed in Table 2.2, indicating the relationship: CLOCK gene – melatonin.

#	PMID	Year	Status	Comment
1	16441555	2006	Seed	"The rhythmic secretion of melatonin by the pineal gland is under control of the circadian clock, conveying the photoperiodic message to the organism."
2	3612591	1987	TN	Neither the title nor the abstract of this paper mention CLOCK gene.
3	7629693	1995	TN	Neither the title nor the abstract of this paper mention CLOCK gene.
4	7949308	1994	TN	Even this paper refers to the interaction of biological clock and melatonin, it does not mention CLOCK gene.
5	7094883	1982	TN	Neither the title nor the abstract of this paper mention CLOCK gene.
6	16687300	2006	TN	Neither the title nor the abstract of this paper refer to melatonin.
7	18045670	2008	FN	"...(2) melatonin is a reliable and stable neuroendocrine output of the circadian clock (versus behavioral output which is sensitive to stress or other factors);..."
8	8836959	1996	TN	Neither the title nor the abstract of this paper mention CLOCK gene.
9	8695891	1996	TN	Even this paper refers to the interaction of biological clock and melatonin, it does not mention CLOCK gene.
10	2979574	1985	TN	Neither the title nor the abstract of this paper mention CLOCK gene.
11	8119657	1994	TN	Neither the abstract nor the title of this paper indicate relationship between CLOCK gene and melatonin.
...				
14	10971628	2000	TP	"The circadian clock in the hypothalamic suprachiasmatic nuclei (SCN) regulates the pattern of melatonin secretion from the pineal gland such that the duration of release reflects the length of the night."
278	12521925	2003	TP	"Melatonin and wheel-running rhythmicity and the effects of acute and chronic light pulses on these rhythms were studied in Clock(Delta19) mutant mice selectively bred to synthesize melatonin."
281	12111535	2002	TP	"Finally, the bases for acute and clock regulation of the key enzyme in melatonin synthesis, arylalkylamine N-acetyltransferase (AA-NAT), are described."
287	19087173	2008	TP	"In mammals, the rhythmic synthesis of melatonin by the pineal gland is tightly controlled by the master clock located in the suprachiasmatic nuclei (SCN)."
294	18662218	2008	TP	"The aim of this work was to investigate the effect of the in vitro circadian-like exposure to melatonin [in the presence or absence of insulin (Ins)] on the metabolism and clock gene expression in adipocytes."
344	11490096	2001	TP	"In mammals, the nocturnal rise in pineal melatonin is regulated by signals from the endogenous clock, the hypothalamic suprachiasmatic nuclei."
463	14972686	2004	TP	"These data suggest that effects of melatonin on clock gene expression are pivotal events in the neuroendocrine response and that pars tuberalis cells can act as molecular calendars, carrying a form of photoperiodic memory."
...				
596				

Table 2.4: Analysis of the top 10 (excluding seed) related citations for a seed paper (in yellow): PMID 16441555, retrieved by Pubmed Related. Papers highlighted in orange are those from the set of 16 papers analyzed in Table 2.2, indicating the relationship: CLOCK gene – melatonin.

hamster (Phodopus sungorus): a tool to study seasonal changes of circadian clock activities."Even from the title it may be concluded that there may be a several similarity topics:

1. papers about hamsters (or even only Djungarian hamsters)
2. papers about CLOCK gene
3. papers about melatonin
4. papers about interactions between CLOCK gene and melatonin
5. papers mentioning all of the above topics together

Currently, PubMed Related retrieves papers using the scenario # 5. The manual analysis of the papers presented in Table 2.4 showed that 8 of top 10 papers refer to studies on circadian phenomenon where investigations were carried out on hamsters (in majority Djungarian hamsters). This is a completely valid outcome in terms of relatedness; however it does not let the user focus on a particular aspect of the paper (such as scenario #4). Only one of the 7 subgraph's papers found in this case is ranked relatively high (position #14), and this is again because the keyword 'hamster' is in the title ("The circadian cycle of mPER clock gene products in the suprachiasmatic nucleus of the siberian hamster encodes both daily and seasonal time.") All the other subgraph's papers, even though they are truly related to the seed paper, are ranked relatively low because they refer to a different animal (mouse – 12521925; avian – 12111535, rats – 19087173, 18662218; rabbits – 11490096; sheep – 14972686). Our methodology did not take into account any of such factors as where the examination took place or type of animal as experiment subject. The interest was only to retrieve papers which very specifically mention any interaction between CLOCK gene and melatonin. Therefore the number of papers in the subgraph (16) is relatively very low comparing to the number of papers retrieved by PubMed Related: 1142 (Table 2.3) and 596 (Table 2.4). We believe that our approach enhances the recent probabilistic retrieval methods. Semantic predications provide the user with the possibility to define which aspect of the seed paper is interesting for him/her and filter out papers which do not refer to this aspect.

Chapter 3

Interoperability and integration in neuroscience

Neuroscience, the field of science of brain and central nervous system, spans over broad spectrum of disciplines. It ranges from biology, chemistry, physics through genetics, physiology, psychology up to statistics and computer science. As a result, neuroscientific research is supported by many different investigators spread all over the world. Those involve not only neurobiologists, radiologists or medical doctors, which produce huge amounts of data, but also journal publishers and reviewers, who enables them to present their findings to the public, or IT specialists and programmers as providers of infrastructure, analysis tools and data storage facilities. Moreover, plenty of centers, consortia and neuro-societies are being set-up in order to organize discussions, workshops and panels to define standards and suggest new movements in neuroscience. Table 3 summarizes most active organizations in this domain.

The most important issues discussed in the domain of neuroscience which should be handled to facilitate more efficient research are data sharing, social services, integration of tools, standard compliance and improvements in paper review system.

Ascoli (2006) emphasized that there are many separated research communities in neuroscience, which refuse to share and exchange experimental data.

Name	Abbreviation	Description
Society for Neuroscience	SfN	Nonprofit membership organization of scientists and physicians who study the brain and nervous system.
International Neuroinformatics Coordinating Facility	INCF	The mission of the INCF is to coordinate and foster international activities in Neuroinformatics.
Neuroscience Peer Review Consortium	NPRC	Alliance of neuroscience journals that have agreed to accept manuscript reviews from other members of the Consortium
International Consortium for Brain Mapping	ICBM	Its primary goal is the continuing development of a probabilistic reference system for the human brain.
Neuroimaging Data Access Group	NIDAG	An international working group dedicated to enhancing access to neuroimaging data in order to advance progress in neuroscience.
International Brain Mapping & Intraoperative Surgical Planning Society	IBMISPS	Non-profit association organized for the purpose of encouraging basic and clinical scientists who are interested or active in areas of Brain Mapping (BM) and Intra-operative Surgical planning (ISP) to share their findings with other physicians and scientists across the disciplines.
Laboratory of Neuro Imaging	LONI	Development of advanced computational algorithms and scientific approaches for the comprehensive and quantitative mapping of brain structure and function.
Alzheimer's Disease Neuroimaging Initiative	ADNI	Its main goal is to determine whether brain imaging can help predict onset and monitor progression of Alzheimer's disease.
Brain Imaging Center	BIC	Cutting-edge research center using a state-of-the-art Positron Emission Tomography (PET) scanner.
Biomedical Informatics Research Network	BIRN	Geographically distributed virtual community of shared resources. Its goal is to advance the diagnosis and treatment of disease.

Table 3.1: A list of major organizations, groups or initiatives devoted to neuroscience with intention to advance progress of research in this domain.

Researchers have expressed concerns that sharing of data can lead to unfair use (Teeters et al., 2008). However, data sharing is important to create trusted collaboration community and is a current topic in debate on future of neuroscience (Liu and Ascoli, 2007) as it is believed that broad data sharing could lead to breakthroughs in our understanding of brain function (Van Horn and Ball, 2008). Invoking online social networks and computer-based communication can support closer relationships and trust (Lampe et al., 2006) hence reducing the resistance to data sharing.

Cheung et al. (2009), and French and Pavlidis (2007) review recent IT approaches trying to resolve two issues in neuroscience: data sharing and integration. The following sections of this chapter will refer to both of them, with special emphasis on text data and literature. In Section 3.1, the general overview of existing biomedical knowledge repositories is presented. Next, Section 3.2 refer to solutions and perspectives in the field of coordinate-based neuroimaging databases (see also Appendices B, F, H). Finally, in Section 3.3 after a review of various existing biomedical literature retrieval systems, I present a novel methodology for context-based literature search in the domain of neuroscience combining biomedical ontologies, natural language processing and statistical ranking approaches (see also Appendix A).

3.1 Biomedical knowledge resources

Neuroscience Information Framework (NIF)[1] (Gardner et al., 2008) is an integrated framework collecting all the data, tools and resources related to the domain of neuroscience, initiated by the U.S. National Institutes of Health in 2005. This very dynamic Web-based repository of neuroscientific information enhances significantly the access to various data sets and analysis tools of the domain.

One of the the major neuroscientific knowledge sources is NeuroLex[2], a semantic Wiki-based lexicon based on the NIF Standard Ontology used for curation of the neuroscience-related terms. It allows to navigate through the terminology of various nervous system levels (e.g. molecular level, cellular level, etc.) and data type categories like drugs, diseases, genes, microarrays, etc.

Another Wiki-based system which records neuroinformatics data from published neuroimaging peer-reviewed papers is Brede Wiki[3] (Nielsen, 2004), a

[1] http://neuinfo.org
[2] http://neurolex.org
[3] http://neuro.imm.dtu.dk/wiki/

direct successor of the coordinate-based neuroimaging database – the Brede database (Nielsen, 2003). Nielsen (2004) discusses both advantages and disadvantages of Wiki services over the ordinary database services. Among others, online versioning and incremental addition of data in Wiki-based services are balanced by the threat of easier data vandalism and more difficult quality control in such systems.

Integration of diverse types of neuroscientific data requires specific scalable approaches and recently Samwald et al. (2010) recalled this issue by describing the SenseLab ontologies as an example of Semantic Web (Berners-Lee et al., 2001) implementation in the domain of neuroscience. In this initiative many databases were semi-automatically translated into machine-understandable ontology file format (OWL) and yet further semantically enriched.

> "The SenseLab ontologies are extensively linked to other biomedical Semantic Web resources, including the Subcellular Anatomy Ontology, Brain Architecture Management System, the Gene Ontology, BIRNLex and UniProt. The SenseLab ontologies have also been mapped to the Basic Formal Ontology and Relation Ontology, which helps ease interoperability with many other existing and future biomedical ontologies for the Semantic Web. (...) The SenseLab ontologies are designed for use on the Semantic Web that enables their integration into a growing collection of biomedical information resources." (Samwald et al., 2010).

Unified Medical Language System (UMLS) developed at the U.S. National Library of Medicine (NLM) "is a repository of biomedical vocabularies developed by the US National Library of Medicine. The UMLS integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. Vocabularies integrated in the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM and the Digital Anatomist Symbolic Knowledge Base. UMLS concepts are not only inter-related, but may also be linked to external resources such as GenBank. In addition to data, the UMLS includes tools for customizing the Metathesaurus (MetamorphoSys), for generating lexical variants of concept names (lvg) and for extracting UMLS concepts from text (MetaMap)" (Bodenreider, 2004). The size and completeness of the UMLS resources make it a very good candidate for support of knowledge and information systems in biomedicine.

Finally, Burns et al. (2008) designed and developed a NeuroScholar[4] system

[4]http://www.neuroscholar.org/

which provides functionalities for efficient management of knowledge from neuroscience literature. Not only it give the access to the primary research literature, but also "it provides a means to add both unstructured and structured annotations to full-text articles as PDF files, an Electronic Laboratory Notebook component and a system to provide support for visualization plugins based on components such as neuroanatomical atlases."

3.2 Coordinate-based neuroimaging databases

Neuroimaging is one of the disciplines in neuroscience referring to various imaging techniques reflecting brain properties such as brain structure and brain functions. Functional localization in brain is normally represented in form of stereotaxic coordinates referring to different brain atlases, e.g. Talairach (Talairach and Tournoux, 1988) or MNI (Evans et al., 1994). Wager et al. (2007, 2009) propose meta-analysis approaches for efficient neuroimaging data summarization and integration, which should facilitate drawing new hypotheses on structure-function correspondence and evaluate and validate consistency and specificity of neuroimaging results from both, human and animal studies. The neuroimaging data sets, thus brain locations represented by coordinates, and papers in which they are recorded are linked together and collected in domain-specific databases, called coordinate-based databases. These are the data sources used to perform above mentioned meta-analysis of the neuroimaging results.

The coordinate-based retrieval systems are designed to search for relevant literature based not only on user's search terms as the ordinary systems do, but also based on brain coordinates as a query. The most popular existing coordinate-based systems in neuroimaging are listed below:

- **Brede database** (Nielsen, 2003) – developed at the Technical University of Denmark and available through the webpage[5] records published neuroimaging experiments that list stereotaxic coordinates in so-called MNI or Talairach space. Presently, close to 4000 coordinates from 186 papers with a total of 586 experiments are available. The data is stored in XML files, and Matlab functions generate static webpages with visualization of the entries in the database. Web-based searching on coordinates is possible from the homepage, but up till now it has required that the researcher manually typed in the query or extracted results from the image analysis program. The Brede database web service provides also links to other neuroscientific resources. While querying the database with a specified

[5] http://neuro.imm.dtu.dk/services/bredeatabase/

coordinate in brain, the user is also able to visualize the location in INC Talairach Atlas. Each publication relates by ID number to other databases like PubMed or BrainMap. Brain regions from each of the experiments are mapped to the services like MeSH, BrainInfo, CoCoMac database or Wikipedia. As the Brede webpages are public, the ordinary Web search engines enable text based search of the Brede database. Furthermore, the researcher may navigate the database via several hyperlinked webpages including brain region, brain function and author ontologies.

- **SumsDB** (Van Essen et al., 2004) – developed in Van Essen Lab at the Washington University in St. Louis and available through the website[6], provides convenient access to various neuroimaging data like surface and volume data, structural and functional data, human and nonhuman primate data, etc. Furthermore, both text-based and spatially based (coordinate-based) searching and data mining is available. Based on the structured user authentication and multiple levels of data access, searches may apply only to public data or to more restricted laboratory data.

- **BrainMap** (Laird et al., 2005) – developed at the University of Texas, online database[7] of functional neuroimaging data storing coordinates in Talairach space. A related software suite consists of three different applications: Sleuth – for database searches and coordinate plotting, GingerALE – for coordinate conversion from MNI space to Talairach and meta-analyses via the activation likelihood estimation (ALE) method, and Scribe – for database entry of published neuroimaging papers together with coordinate data sets.

- **AMAT**[8] – a meta-analysis toolbox for the SPM neuroimaging environment developed by Antonia Hamilton at the University of Nottingham provides coordinate-based search for over 5000 coordinates from 213 published papers of which some were derived from the Brede Database. The coordinates are in MNI or Talairach space. The toolbox can locate neighboring coordinates to a given coordinate, as well as publications for a given author or year.

Another system, BrainKnowledge (Hsiao et al., 2010), fits to the field of neuroimaging, but instead of extracting coordinates from papers, detects and stores brain anatomical structures from abstracts of papers. It is Java-based system associating fMRI data-analysis and literature search functions. Three major features of BrainKnowledge are: searching for brain activation models by function, searching for function by brain structure, and comparing user's fMRI experimental results with already published studies. Recently, the system uses a database

[6] http://sumsdb.wustl.edu/sums/
[7] http://brainmap.org/
[8] http://www.antoniahamilton.com/amat.html

which consists of 15,413 abstracts from 1032 journals dating from 1992 to the present.

A drawback of such systems like BrainKnowledge is the fact that it is a separate, standalone program which is not integrated with any common neuroimaging environment like Statistical Parametric Mapping (SPM) (Friston et al., 1994) or FMRIB Software Library (FSL) (Smith et al., 2004b). Additional work is required from researchers to import/export functional experimental results from one environment to another in order to perform complementary analyses or search for relevant neuroimaging literature.

AMAT toolbox for SPM, discussed above, is one of the responses for the issue of tool integration with common neuroimaging environments. We propose a BredeQuery plugin for SPM environment which links SPM to the coordinate-based search engine in the Brede database . BredeQuery is able to 'grab' brain location coordinates from the SPM windows and enter them as a query for the Brede database. Moreover, results of the query can be displayed in a MATLAB window and/or exported directly to some popular bibliographic file formats like BibTeX or Reference Manager. The detailed description of the BredeQuery plugin for SPM is given in Appendix B.

To extend further the results pool of the small Brede database in the plugin and offer to the user an opportunity to retrieve more recent and relevant papers, there is a need for a direct integration of the current tool with a larger and more comprehensive biomedical literature database like the PubMed database (see Appendix H). A methodology which extends the BredeQuery plugin in such a way and besides is a proposal of a novel approach for context-dependent semantic literature retrieval. It is included in the recent version of the BredeQuery plugin, see Figure 3.1, and is further discussed in Section 3.3.

3.3 Biomedical literature retrieval

Existence of such knowledge resources as described in Section 3.1 enables developers to build various literature retrieval systems. Herein, the overview of such systems in neuroscience and biomedicine is given.

NeuroText (Crasto et al., 2003) and NeuroExtract (Crasto et al., 2007) systems aim respectively in populating neuroscience databases, and performing the integrated retrieval of Internet-based information relevant to neurosciences. NeuroText was developed for text-mining of abstracts from neuroscience journal articles in order to identify relevant domain keywords which allow further clas-

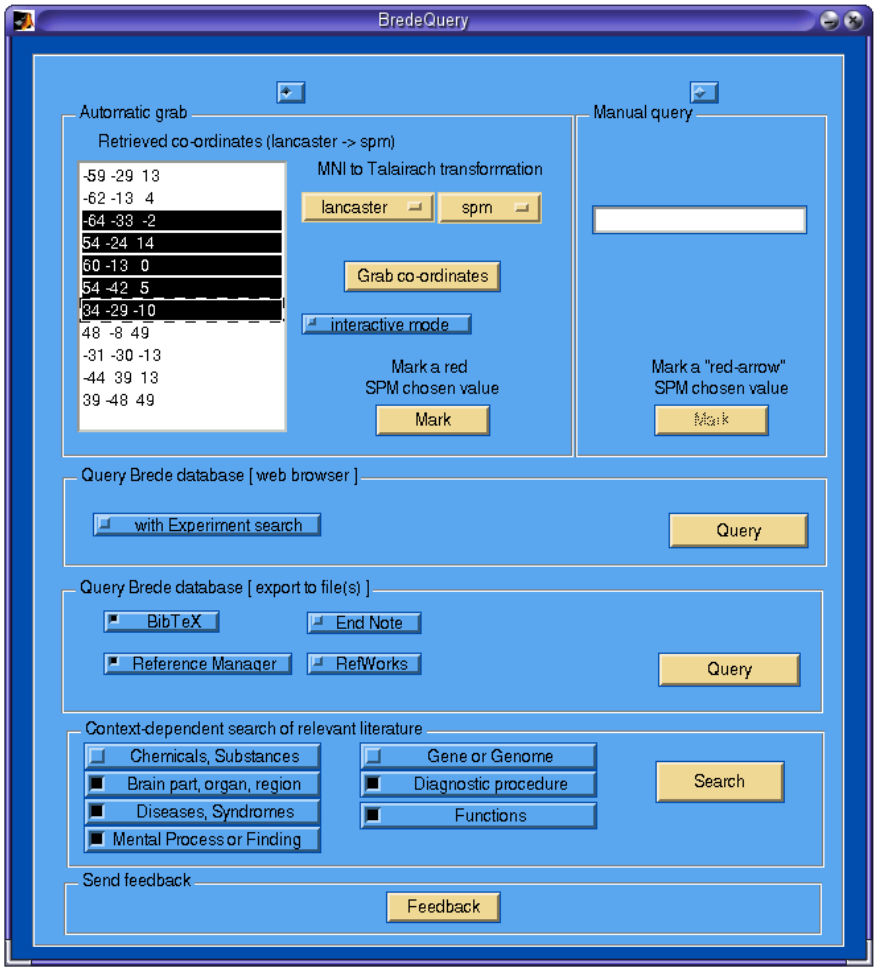


Figure 3.1: Graphical user interface of the BredeQuery plugin for SPM environment. Brain activation coordinates grabbed from the SPM results window are used for further search of relevant literature. Beforehand, coordinates may be transformed from MNI to Talairach atlas through Brett (Brett, 1999) and Lancaster (Lancaster et al., 2007) transformations. Relevant papers are retrieved from the Brede database. Biomedical concepts extracted from the title and abstract of highest ranked paper are classified into one of seven semantic categories. User is able to select categories of interest, thus define a context of search, which allow to retrieve other relevant publications from the PubMed database (Appendix A).

sifications. It facilitates creation of bibliography databases for a specific domain of interest.

NeuroExtract system retrieves neuroscience related information from genomic and proteomic repositories: SwissProt, Gene Expression Omnius and Protein Data Bank, which are not primarily considered as neuroscience knowledge sources. NeuroExtract allows simultanous querying of these various data sources, what significantly enhances interoperability and facilitates the users. NeuroExtract facilitates also easy adaptation of additional sources.

Textpresso for Neuroscience (Müller et al., 2008) is another text mining system which performs literature search in the domain of the neuroscience. In this search engine, users specify not only search keywords, but also one or a combination of categories in which the search should be performed. It is presented that thanks to the definition of categories, a keyword query is more refined and meaningful. The functionality of Textpresso for Neuroscience bases not only on abstracts but also on full texts of papers, which are collected in local database.

In terms of general biomedical literature retrieval, the PubMed online service "comprises over 20 million citations for biomedical literature from MEDLINE, life science journals, and online books."[9] Apart from simple text search, Pubmed offers the opportunity to use Medical Subject Headings (MeSH) to define context of search and retrieve fewer, but more relevant papers. Any biomedical paper found in PubMed is indexed with a few MeSH headings, manually assigned by human indexers employed at U.S. National Library of Medicine, which represent the most important topics of a given paper. MeSH terms are organized in a parent-child taxonomy, e.g. a MeSH heading 'Leukemia' is a type of neoplasm therefore it is a child of the MeSH heading 'Neoplasms'. On the other hand there are 9 different types of leukemia, e.g. 'Enzootic Bovine Leukosis' which are defined as children of 'Leukemia'.

For demonstration purposes, let's assume we are interested in finding papers about leukemia. A simple PubMed query with a term *leukemia* returns 235542 publications. Next, if we specify that we are just interested in papers which are indexed with a MeSH heading 'leukemia' (*leukemia[mh]*), Pubmed returns 178456 papers, which means that over 20% of papers found in the simple query were eliminated because even they mention a word 'leukemia', this is not one of their main topics. Each of the 178456 papers is indexed with 'Leukemia' MeSH or any of its children. Finally, if we want to exclude children from a search, a query *leukemia[mh:noexp]* is yet more selective and returns 'only' 45879 papers. Construction of more complex PubMed queries employing MeSH headings and AND/OR/NOT logical operators facilitate the users in finding fewer papers but

[9] http://www.ncbi.nlm.nih.gov/pubmed/

more relevant and related to their interests.

Even the MeSH taxonomy is a very simple example of biomedical ontology where all concepts are related to each other only through a single, IS A, relationship, it definitely shows the capability of ontologies in enhancement of literature retrieval. Real biomedical ontologies consist of many more relationships through which various biomedical concepts may connect to each other, e.g. drugs cure diseases (relationship TREATS), body substances have influence on each other (relationships INHIBITS, STIMULATES), relative position of body organs to each other (relationships PART OF, LOCATION OF). Furthermore, any concept in an ontology has one or more semantic type defined describing its meaning. Drug, disease, body substance, body part are examples of semantic types assigned to biomedical concepts. Consequently, employment of well-constructed domain ontologies should be considered as a potentially very powerful step towards improvement of recent literature retrieval approaches.

3.4 Context-dependent literature retrieval

Based on the keyword classification and categorization ideas from NeuroText, NeuroExtract and Textpresso of Neuroscience and the potential of biomedical ontologies in the development of literature retrieval, we propose a prototype of methodology for context-dependent literature retrieval in neuroscience (later called here as CDLR). The objective of this system was to support neuroimaging literature retrieval through the integration with BredeQuery plugin. Recently, this plugin takes a brain activation volume (set of coordinates) as an input, later mapped to a neuroimaging paper, reporting the most similar brain volume, returned by the Brede database. The abstract and title of this paper is further used as an input to our CDLR methodology.

The CDLR's input, any text related to the domain of neuroscience such as paper's title and abstract, is first analyzed by the natural language processing and ontology mapping software Metamap (Aronson, 2001), which maps biomedical terms to the concepts from UMLS ontology. Later, these concepts are carefully filtered and classified into seven semantic categories earlier defined to fit the domain of neuroscience. Later, user defines context of search by selecting relevant semantic categories of interest. All the keywords belonging to the selected categories are joined together to a logical query executed on PubMed database, which returns a list of related papers. The details of the CDLR methodology is provided in Appendix A.

Section 3.4.1 presents the details of further development of the ranking system

for the presented CDLR methodology, not included in Appendix A. Finally, Section 3.4.2 discusses results of a pilot study on the implemented methodology.

3.4.1 Paper ranking with semantic predications

The type of papers returned by the CDLR system depends on which of the seven semantic categories (see Table 1 in Appendix A) are chosen by the user, specifying a context of search. The returned papers are later ranked by relevance to the input and the context of search. For paper scoring and ranking purposes we decided to employ semantic predications extracted by the SemRep system, presented in Chapter 2 (Section 2.1).

First, semantic predications of the input paper and the output papers are retrieved from the SemRep's database (Section 2.1.1). Every predication of any output paper is scored by comparing it with every predication of the input paper. Finally, a sum of all predication scores reflects a final paper's score. For example, the input paper A, consists of a set i predications: $\{P_1, P_2, \ldots, P_i\}$. User has also chosen the semantic categories of interest (context of search), $C = \{C1, \ldots, C7\}$. The chosen categories get the value of 1 while the rest has the value of 0 assigned. There is also a list of j retrieved output papers $\{A_1, A_2, \ldots, A_j\}$, each having a number of predications. A final score ($finalscore$) for a random output paper A_k $(1 < k < j)$ having m associated predications $\{Pk_0, Pk_1, \ldots, Pk_m\}$ is calculated as

$$finalscore_k = \sum_{n=1}^{m} \sum_{p=1}^{i} pscore(Pk_n, P_p) \cdot cscore(C, Pk_n). \tag{3.1}$$

According to Figure 2.2 in Chapter 2, two semantic predications, P_1 and P_2, have a triple form and consist of respective subjects (S_1, S_2), predicates (Pr_1, Pr_2) and objects (O_1, O_2). An auxiliary function $pscore(P_1, P_2)$ compares this pair of predications and returns a score,

$$pscore(P_1, P_2) = \begin{cases} 5, & \text{if } S_1 = S_2 \wedge Pr_1 = Pr_2 \wedge O_1 = O_2; \\ 3, & \text{if } (S_1 = O_2 \wedge Pr_1 = Pr_2 \wedge O_1 = S_2) \vee (S_1 = S_2 \wedge O_1 = O_2); \\ 2, & \text{if } (S_1 = S_2 \vee O_1 = O_2 \vee S_1 = O_2 \vee O_1 = S_2) \wedge Pr_1 = Pr_2; \\ 1, & \text{if } S_1 = S_2 \vee O_1 = O_2 \vee S_1 = O_2 \vee O_1 = S_2; \\ 0, & \text{otherwise;} \end{cases} . \tag{3.2}$$

Finally, function $cscore(C, P)$ checks if the semantic categories assigned to subject and object of a predication P are chosen by the user in the search context

C. If both, subject and object's semantic categories are chosen in context C, function *cscore* returns 1. If only one, subject or object's semantic category is chosen in context C, function *cscore* returns 0.5. Otherwise, *cscore* returns 0. The objective of this function is to determine if a given predication is completely out of scope of user's search context and should be disregarded or if it converges with the context and should be included in the final score calculation.

3.4.2 Preliminary results

This section presents results from a pilot study evaluating the CDLR methodology. The objective was to check if the papers returned by CDLR methodology report similar brain activation volumes (sets of coordinates) comparing to the volumes found in the CDLR's input paper (seed). The context of search defined for the CDLR methodology in this analysis includes just two semantic categories. We were interested in finding papers similar to the seed paper by location in brain (chosen category: *Body (Brain) Part, Organ or Region*), which describe the neuroimaging experiments using the functional magnetic resonance imaging technique (fMRI) (chosen category: *Diagnostic procedure*).

First, a full text of the seed was manually examined and all brain activation volumes were extracted. Then, for each volume the Brede database was queried and retrieved a set of top 20 papers, which report activation volumes in the most similar brain locations. Each of the 20 papers returned by the Brede database has a similarity value calculated. In Brede database a Gaussian/Euclidean form for calculating the similarity between two volumes is used (Nielsen and Hansen, 2004):

$$s_{q,e} = \frac{1}{\sqrt{N}} \sum_{m=1}^{M} \sum_{n=1}^{N} \exp\left(-\frac{(x_{m,q} - x_{n,e})^2 + (y_{m,q} - y_{n,e})^2 + (z_{m,q} - z_{n,e})^2}{2\sigma^2} \right), \quad (3.3)$$

where σ is set to 10 millimeters, $(x_{m,q}, y_{m,q}, z_{m,q})$ is the mth of M three-dimensional query coordinates, while $(x_{n,e}, y_{n,e}, z_{n,e})$ are the nth of N three-dimensional coordinates in the Brede database.

In this analysis, we have recorded the similarity value of the last, the twentieth, paper found on the list of top 20 similar papers. This value was further used by us as a threshold.

Later, the detailed investigation on the papers returned by the CDLR methodology was performed. Starting from the top of the CDLR's list of papers ranked using semantic predications (Section 3.4.1), we have analyzed them manually, one-by-one, in order to find these which report in the full text any brain activation volumes in form of coordinates. Since many of the papers did not contain

Seed	CDLR				Pubmed Related (PR)			
PMID	#	PMID	Similar	In PR	#	PMID	Similar	In CDLR
18775476	3	15699291	•	-	3	19347877	•	#24
	4	19556348	•	#20	5	15953488	•	-
	5	16339042	-	-	6	18457504	•	#88
15876491	7	20027574	-	#81	4	18684074	-	#41
	8	20171291	-	-	5	19061937	-	#170
	10	13679404	-	-	6	18571795	-	-
17881514	2	15528086	•	-	5	16167193	•	-
	4	15528081	-	#23	6	18524903	-	-
	5	11969314	•	-	7	10725932	•	-
17103153	2	15885507	•	-	2	14622239	-	#19
	4	16421332	•	-	3	16839610	-	-
	5	15958738	-	-	4	11369672	•	#1632
15050584	2	19965853	•	-	2	15050567	•	-
	6	17275336	•	-	5	14980562	•	-
	22	15580514	•	-	7	16757183	•	-

Table 3.2: The results for a pilot study evaluating the performance of the context-dependent literature retrieval (CDLR) methodology. The analysis was carried out for five different input papers (seeds). The results for both, CDLR methodology and Pubmed Related are presented. Column [#] shows the position of the paper on the ranked paper list. Column [PMID] shows the paper identification number in Pubmed database. Column [Similar] displays a symbol (•) if the similarity value, measured between at least one brain activation volume from a given paper and volumes reported in the seed paper, is above the threshold. Finally, columns [In PR] and [In CDLR] show the position of a given paper in the ranked list retrieved by Pubmed Related and CDRL, respectively.

any coordinates and the time frame available for this analysis was limited, we continued until three papers with the reported brain activation volumes were found.

Finally, using Equation 3.3, a similarity was calculated between the seed's volumes and each volume of the three selected papers. If at least one volume from a selected paper had a similarity value higher than previously calculated threshold, it was marked as a paper similar to the seed.

We have pursued the analysis for five different seeds. In parallel, the papers returned by the Pubmed Related feature were investigated in the same manner. The results are shown in Table 3.2. In four of five presented cases there was, for both CDLR and Pubmed Related, one or more highly ranked papers similar, by brain activation location, to the seed paper. Moreover, CDLR and Pubmed Related seem to perform in this study in a complementary manner since the highly ranked, similar to seed, papers returned by CDLR do not overlap with papers, similar to seed, returned by Pubmed Related. Consequently, it may be

assumed here that the execution of this two methods in parallel in a common web service might bring an useful enhancement to the process of retrieval of neuroimaging studies.

Chapter 4

Conclusion

In this thesis we have discussed a number of semantic approaches for literature retrieval and knowledge discovery applied to the domain of biomedicine. We showed that semantic predications extracted from the biomedical texts during the process of semantic natural language processing can lead to powerful tools enhancing the probabilistic and statistical methods implemented recently in the literature retrieval systems.

This thesis aimed to provide a compact view on the background and principles of the covered topics and summarize the developed methodologies, presented and discussed in full in their respective appendices. In Section 2.1 of Chapter 2 we introduced the notion of semantic predications and discussed the current semantic natural language processing systems which specialize in extraction of such predications from biomedical texts. Further sections of Chapter 2 introduce the field of literature-based discovery (LBD) and discuss its pioneer and recent methodologies. Finally, in Section 2.5 we present the discovery browsing methodology (Appendix C), the extension of current LBD approaches, which employs semantic predications and fundamental measures of graph theory to search through the huge biomedical literature repositories for previously unknown or poorly studied knowledge. The main thrust of Chapter 3 is manipulation and retrieval of data in the domain of neuroimaging, the subfield of neuroscience. The detailed review of the current knowledge repositories and databases in neuroimaging is followed by a discussion on the literature retrieval

in neuroscience. We presented in Section 3.4 the novel methodology for retrieval of neuroimaging studies in a context-dependent manner (Appendix A) and its integration with the common neuroimaging environment, Statistical Parametric Mapping, through the BredeQuery plugin (Appendix B) which provides an interface for coordinate-based literature searching.

Discussion and Future work

We discuss here the outcomes from the pilot evaluations of the developed methodologies and propose their further development directions.

Literature-based discovery

The discovery browsing methodology presented in Appendix C and the analysis results given in the final section of Chapter 2 demonstrate the utility and power of semantic predications in literature manipulation. The results from the depressive disorder study, where the poorly studied interaction of circadian phenomena, inflammation and the neurotransmitter norepinephrine in depression was highlighted, were very positively discussed and evaluated in Appendix C by three domain experts. The advantage of the discovery browsing methodology and underlying path analysis is the fact that it provides an user interactive, principled way of navigating through selected aspects of some biomedical domain extracted from a MEDLINE database, a huge biomedical literature repository. The analysis of the final results of the depressive disorder study, presented in Section 2.5.3, demonstrates that our methodology allows the user to navigate through literature from diverse biomedical domains, some of which may go beyond the normal scope of interest of the user, which consequently may lead to new hypotheses and discovery of new knowledge. The presented methodology is recently employed in two ongoing studies (sleep apnea and restless legs syndrome) carried out at the U.S. National Library of Medicine. The current results obtained in the sleep apnea study using the discovery browsing methodology allowed to build a hypothesis on a potential pharmacological therapy for this disorder including a combination of acetylcholine agonist, glutamate agonist and gamma-Aminobutyric acid antagonist.

We plan to pursue with further development of the discovery browsing methodology which includes the design of various path rankings, e.g. combining simple and weighted degree centrality and checking strong/weak predications in the paths. In the recently implemented rankings, we do not take into account how

many papers contributed to a given A – B predication. The more papers reveal a given predication, the stronger it is. We are planning to include this measure in the design of future ranking approaches. In addition, even semantic predications represent directed relationships between two given concepts (A – *predicate*→ B and A ←*predicate*– B are two different predications), for simplicity purposes our methodology does not rely recently on the predication direction. Nonetheless, we are planning to take advantage of this knowledge in the future. Finally, since the presented methodology uses semantic predications given by the semantic natural language processing systems, like the SemRep system, thus, consequently, further improvements of such systems and the definition of new predicates for the domains recently not covered by SemRep, e.g. public health, would be very beneficial.

The flexibility of semantic systems and methodologies presented in this thesis results in a very broad spectrum of potential applications. Apart from the literature retrieval and literature-based discovery, which proved to link various knowledge domains, semantic approaches like 'discovery browsing' methodology (Appendix C) might be beneficial for educational purposes (summarization of knowledge) as well as for innovation- and novelty-oriented analysis of biomedicine related documents and applications.

Literature retrieval in neuroimaging

The context-dependent literature retrieval (CDLR) methodology for neuroscience presented in Appendix A retrieves similar papers given the input and context of search. The results of a pilot study presented in Section 3.4.2 demonstrate that our methodology performs well while searching for neuroimaging papers referring to similar locations of brain activations. It confirms the potential of the CDLR methodology to extend results pool of much smaller coordinate-based neuroimaging databases. We have integrated the BredeQuery plugin (Appendix B) with the CDLR methodology to facilitate extensive literature searching directly from the neuroimaging environment. Furthermore, the positive results from the pilot study affirm also the novel way of paper ranking with the use of semantic predications (Section 3.4.1) and outline the potential of semantic predications for improvements of the current literature retrieval algorithms (also discussed in Section 2.5.4, Chapter 2).

We are planning to carry out in the near future a more extensive evaluation based on this pilot study for a bigger set of input papers. Moreover, the performance of the BredeQuery plugin with CDLR methodology is to be verified and evaluated by the domain experts. To conclude, in the era of extremely high growth rate of published biomedical studies, we have demonstrated that a combination of

semantic approaches, which can efficiently filter out papers for very specific biomedical topic, and the recently very popular probabilistic and statistical retrieval systems may facilitate users in navigating through a huge amount of papers and finding the relevant ones.

Appendix A

Context-dependent literature search: a support for functional imaging coordinate-based databases

---

Bartłomiej Wilkowski, Lars Kai Hansen. Context-dependent literature search: a support for functional imaging coordinate-based databases. *BMC Neuroscience*, 2011, (pp. 20). Submitted.

Context-dependent literature search: a support for functional imaging coordinate-based databases

Bartłomiej Wilkowski*[1,2], Lars Kai Hansen[1,2]

[1]Technical University of Denmark, DTU Informatics, Richard Petersens Plads, B321, DK-2800, Kongens Lyngby, Denmark
[2]Center for Integrated Molecular Brain Imaging – Cimbi.org

Email: Bartłomiej Wilkowski*- bw@imm.dtu.dk; Lars Kai Hansen - lkh@imm.dtu.dk;

*Corresponding author

Abstract

Background: The growth of published neuro-imaging articles and experimental results in neuroscience, brings a demand for dedicated information retrieval tools for this specific biomedical domain. There is a need to expand results pool of the functional imaging coordinate-based databases like SumsDB, BrainMap or Brede, because, despite their limited size, they are arguably richer for neuroimaging than conventional keyword-based retrieval services.

Results: In this work we propose a general methodology for linking coordinate-based and keyword-based retrieval systems. The input to the presented pipeline of methods is any paper retrieved by a coordinate-based service. First, an automatic extraction of significant keywords is performed by mapping noun phrases to Unified Medical Language System's ontological concepts using the Metamap software. Each of the extracted keywords is automatically classified into one or more semantic groups relevant to neuroscience. The semantic groups and, associated with them, extracted keywords, are later used for construction of logical queries executed on the PubMed database. The context of the search is defined by selection of semantic groups of interest. A semantic group representing brain parts is always stipulated in any search to ensure that the retrieved papers are similar to the input article by brain part or location.

Conclusions: The discussed methodology may be considered as an extension, in the domain of neuroscience, of the common publication search engines like PubMed or Google Scholar. We propose integration of the methodology with the BredeQuery plugin for enabling searches directly from within the Statistical Parametric

Mapping, a very popular environment in function neuroimaging. Finally, the flexibility of the approach presented in this work allows adjustment of the methodology also to other biomedical domains.

Background

Breakthroughs in functional brain imaging technology and an increased interest in neuro-psychology have stimulated an explosive growth of published neuro-imaging papers and experimental results. As in other expansive areas of medicine there is a strong need for efficient literature and knowledge retrieval from the many on-line databases and other data sources.

There are already several tools available to assist the neuroimaging scientist. Comprehensive literature databases provide users with a common web interface where the papers can be retrieved based on keywords. The most popular tool of that kind is the PubMed database. It keeps references to over 19 million biomedical publications including entries from MEDLINE database (medical and health journals), and life science journals. Also keyword based generic science search engine Google Scholar is extremely helpful for the neuroimaging researcher.

In neuroimaging most published results take the form of activation locations associated with given behavior, and this has given rise to specialized location databases. The major location/coordinate-based databases are: SumsDB [1], BrainMap [2] and Brede [3]. The main challenge for this kind of database is the labor intensive data entry process which has limited the coverage and resulted in limited use.

In spite of the limited size of functional imaging coordinate-based databases, their functionality and methods for paper retrieval are arguably richer for neuroimaging than ordinary, keyword-based search services.

In addition, coordinate-based search is possible also from from the brain imaging analysis pipeline Statistical Parametric Mapping (SPM). The latter functionality is implemented by the BredeQuery plugin [4]. It is an example of an SPM extension where user is able to grab activation coordinates from

SPM results window and perform querying of the Brede database for related literature.

To overcome the limitated coverage of pure coordinate databases, and draw from the large pools of information present on-line new tools are needed. Therefore, we here present a general methodology aimed at expanding results from the coordinate databases into the wider context of neuroscience.

Related work

NeuroText [5] and NeuroExtract [6] systems aim respectively in populating neuroscience databases, and performing the integrated retrieval of Internet-based information relevant to neurosciences. NeuroText was developed for text-mining of abstracts from neuroscience journal articles in order to identify relevant domain keywords which allow further classifications. It facilitates creation of bibliography databases for a specific domain of interest.

NeuroExtract system retrieves neuroscience related information from genomic and proteomic repositories: SwissProt, Gene Expression Omnius and Protein Data Bank, which are not primarily considered as neuroscience knowledge sources. NeuroExtract allows simultanous querying of these various data sources, what significantly enhances interoperability and facilitates the users. It is flexible for further adaptation of additional sources.

The notion of keyword identification and relevant information extraction is also considered in our methodology. The extracted keywords are further classified into semantic groups, which enable the definition of a context in which search is to be performed. The main aim of our approach is to provide an interoperability between structured datasets containing volumes of brain location coordinates and the bibliographic databases.

Textpresso for Neuroscience [7] is another text mining system which performs literature search in the domain of the neuroscience. In this search engine, users specify not only search keywords, but also one or a combination of categories in which the search should be performed. It is presented that thanks to the definition of categories, a keyword query is more refined and meaningful. The functionality of Textpresso for Neuroscience bases not only on abstracts but also on full texts of papers, which are collected in local database.

One noticable difference in the design of Textpresso for Neuroscience and the presented methodology is the

fact that in our approach we do not expect keywords as an input for searching, but any arbitrary English text. It could be an abstract of any paper or author's own manuscript. Keywords are extracted automatically using the Metamap software and an additional set of procedures. Still, both in Textpresso for Neuroscience and our methodology, categories (semantic groups) are to be defined for the search, what allows to narrow down or broaden the context from which related papers are to be retrieved. Finally, in contrary to Textpresso for Neuroscience, we do not create our own database of publications, but logical queries are executed on one of the biggest and most popular medical database, the PubMed database, which ensures that most relevant papers, including historical and very recent work, are included in the search's retrieved related papers list.

There are numerous research activities at the National Institutes of Health based on the Unified Medical Language System (UMLS) related to the areas like ontological development, knowledge extraction, etc. One example is SemRep [8] system, which is able to recover semantic predications from biomedical texts with help of partial syntactic analysis derived by SPECIALIST Lexicon [9] and part-of-speech tagging using the MedPost tagger [10].

In addition, Semantic MEDLINE is a recent summarization initiative – a web-service, which manages Pubmed's searches, provides visualizations of the semantic predications extracted from MEDLINE citations, later linked to various structured resources [11]. Functionality of the Semantic MEDLINE relies on both the SemRep and MetaMap [12] systems.

MGREP, developed at the University of Michigan [13], is a tool similar in functionality to the NIH's Metamap software. Both of them are tools which map a natural language text to ontological concepts. A comparison of the two systems was carried out recently in [14] where among others the flexibility, extensibility and speed were discussed. The advantages of MGREP over Metamap is the scalability across the dictionaries used and execution speed. Metamap is tightly coupled with UMLS and it is difficult to introduce external data sources. However, it does map natural language text to UMLS concepts efficiently. Choosing MGREP or Metamap thus depends on the application.

Finally, our aims are in part shared with the Open Biomedical Annotator developed at the Stanford University. It is a web-service which annotates public datasets, based on the textual metadata, with biomedical ontologies [15]. It currently uses MGREP for concept mapping and the data sources used are:

UMLS and NCBO BioPortal ontologies.

Methods

Main components of the methodology

The key components of the the presented methodology are the Metamap software [12], which enables mapping of biomedical texts phrases to UMLS ontological concepts, and the Brede database [3], which records neuroimaging publications together with brain activation coordinates.

Metamap software and UMLS

Metamap is developed by Dr. Alan Aronson at the National Library of Medicine (NLM) and uses the UMLS terminology and resources.

UMLS [16, 17] is a data source widely used in data repositories for biomedical applications and computer systems and consists of:

1. medical vocabulary database (Metathesaurus) – contains over one million biomedical concepts taken from over 100 source vocabularies.
2. Semantic Network – 133 broad categories and fifty-four relationships between categories for labeling the biomedical domain. It is used in applications to help interpret meaning.
3. SPECIALIST Lexicon and Lexical Tools – lexical information and programs for language processing.

The purpose of the Unified Medical Language System (UMLS) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health[1].

Metamap is a natural language processing and computational linguistics based software, which maps simple noun phrases to UMLS Metathesaurus concepts. The natural language processing is performed using MedPost tagger, which achieves over 97% accuracy on MEDLINE citations [10]. We use a 2009 Metamap version, in which the processing speed and program functionalities were improved significantly comparing to the previous releases. Metamap offers, apart from an ordinary raw output, XML output, where metadata about the introduced text phrases is presented: part of speech, mapped concepts identifiers and labels, preferred concepts, semantic types and knowledge sources for mapped concepts, etc. Moreover, Metamap can be easily configured to the user's needs by simple selection of relevant parameters.

[1] http://www.nlm.nih.gov/research/umls/new_users/online_learning/index.htm

One of Metamap's options currently used in our methodology is Word Sense Disambiguation (WSD), accuracy of which was enhanced in Metamap 2009 version by increasing the number of algorithms and introducing voting mechanism for final ambiguity resolution [18].

Brede database

The Brede database [3] is a web-service which records published neuroimaging experiments that list stereotaxic coordinates in so-called MNI or Talairach space [19]. Presently, close to 4000 coordinates from 186 papers with a total of 586 experiments are available. Apart from basic data about the neuroimaging studies, the Brede database stores also links to many other neuroscientific resources like PubMed, BrainMap, CoCoMac, SumsDB or Wikipedia. Furthermore, it provides graphical visualizations of the experiment's data sets. In spite of the fact that the presented database is very small, it contains scientific data on almost all brain regions and functions, and that is why it can be considered as a great mediating knowledge source for other, bigger neuroinformatic initiatives, which require well-defined neuroscientific reference data and information.

Methodology description

The discussed methodology is a pipeline of several steps, where the input is an arbitrary English neuroscientific text (here retrieved by a coordinate-based retrieval system) and the output is a list of related publications – see Figure 1.

Definition of semantic groups for neuroscience

Since UMLS consists of various medical vocabularies for all areas of medicine and biomedicine and, as mentioned, our methodology is currently aimed at neuroscience, thus it was necessary to define which semantic types, present in UMLS, are not relevant for neuroscience. For that purpose, we used all abstracts and titles of articles (186) stored in the Brede database, which were further analyzed by Metamap. After the classification of the mapped concepts by UMLS's semantic types, we filtered out the irrelevant semantic types, therefore these which do not carry any essential information about neuroscience field.

The outcome of the above described procedure is that, from 135 available semantic types available in UMLS, a final set of semantic types for neuroscience consists of 30, later called "valid", semantic types (105 filtered out). This assures that in further processing of any neuroscience text only valuable,

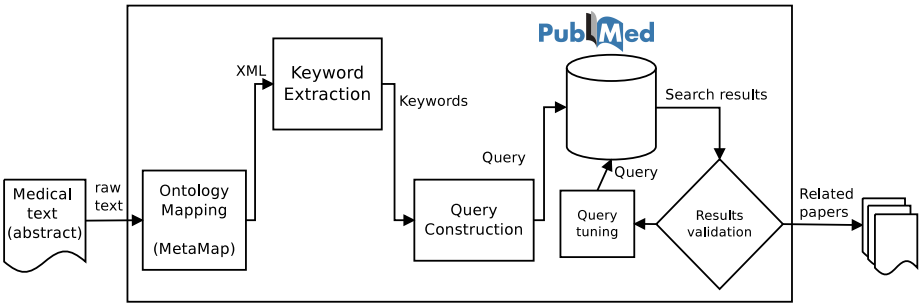


Figure 1: A pipeline of steps representing the methodology. A neuroscientific text (input) is the input. It is later analyzed by the Metamap software, which outputs an XML file with the results (ontology concepts and mappings, part-of-speech tags, scores for mapped concepts). The data from the XML file is retrieved and keyword extraction is carried out. The extracted keywords are semantically grouped and then used in a query construction. The query is sent to PubMed database, which returns a set of related publications (output). The Results validation and Query tuning steps are two proposal phases (not yet tested) and could be used if the modification of the previously constructed query is needed, e.g. if PubMed did not return any papers or returned too many of them.

informative keywords of the field will be selected (concepts with "valid" semantic types).

Nevertheless, some of the 30 selected semantic types still reference to the same aspects or topics of neuroscience, thus consequently, these were grouped together in seven bigger and more descriptive semantic groups (categories), see Table 1. These semantic groups are used for classification of extracted keywords.

Ontological mapping by MetaMap

The first step of the methodology is to process an input text by MetaMap. As it was mentioned before, MetaMap performs NLP (natural language processing) on the plain text and, afterwards, it searches for the most relevant ontological concepts in UMLS Metathesaurus. This process is often called as ontological mapping of raw text and it should be understood as finding the most relevant medicine related standardized dictionary concept for each noun phrase from a given text input.

During the text analysis, MetaMap divides a whole text into utterances, which are later split into noun phrases. Firstly, each noun phrase is part-of-speech tagged. Later, the software tries to find concepts in the UMLS ontology (ontological candidates) for every noun phrase. Each candidate obtains a score. There can be multiple candidates for a single noun phrase. Consequently, based on scores, the software makes a

Semantic group	Semantic types
Body (Brain) Part, Organ or Region	Body Location or Region (blor), Body Part, Organ, or Organ Component (bpoc), Body Space or Junction (bsoj), Cell Component (celc)
Disease or Syndrome	Disease or Syndrome (dsyn), Neoplastic Process (neop), Mental or Behavioral Dysfunction (mobd)
Functions	Organism Function (orgf), Organ or Tissue Function (ortf), Molecular Function (moft), Pathologic Function(patf), Physiologic Function (phsf)
Mental Process or Finding	Mental Process (menp), Finding (fndg), Individual Behavior (inbe), Natural Phenomenon or Process (npop), Sign or Symptom (sosy)
Diagnostic procedure	Diagnostic procedure (diap), Laboratory or Test Result (lbtr)
Chemical or Substance	Amino Acid, Peptide, or Protein (aapp), Biologically Active Substance (bacs), Lipid (lipd), Neuroreactive Substance or Biogenic Amine (nsba), Nucleic Acid, Nucleoside, or Nucleotide (nnon), Organic Chemical (orch), Pharmacologic Substance (phsu), Receptor (rcpt), Vitamin (vita), Hormone (horm)
Gene or Genome	Gene or Genome (gngm)

Table 1: 30 carefully selected semantic types encapsulated in seven categories (semantic groups) which represent various contexts of search used by the presented methodology.

decision on which candidate is the best mapping for a given noun phrase.

Multiple mappings for the same noun phrase appear, because natural language is highly ambiguous. In order to obtain a final, the best and only one mapping, the Word Sense Disambiguation (WSD) engine is available in Metamap. Since in the release of Metamap from 2009 the accuracy of WSD has increased significantly, therefore we have decided to employ it in the current approach.

From MetaMap we can also derive the properties of each mapping concept: semantic types and knowledge sources. Each concept in UMLS Metathesaurus is assigned to one or more semantic types. There exist also certain relationships between various semantic types which create the UMLS Semantic Network. Currently,

there are 135 semantic types defined in UMLS.

From among a few output formats offered in MetaMap, a machine-understandable XML (Extensible Markup Language) is to be used in our methodology.

Semantic keyword extraction

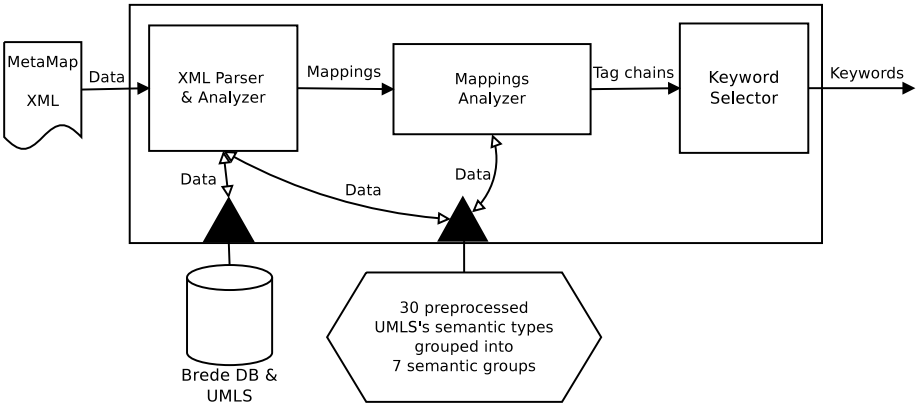


Figure 2: Semantic keyword extraction step. First, the Metamap XML output is parsed and analyzed (ontology mappings are retrieved). Then all the mappings are classified and represented as a chains of specific tags. Finally, the tag chains are processed and the not relevant ones (these which carry inappropriate semantic type or part of speech) are filtered out. The remaining mappings are returned as keywords.

This section describes how keywords are extracted (see Figure 2) given the Metamap's XML output produced in the previous step.

The first step performed during the keyword extraction step is parsing and analyzing the Metamap's XML output file. All the possible mappings given by Metamap need to be analyzed and only relevant ones can be selected for further analysis. Every mapping consist of one or more concepts, while every concept is created from one or more words. The choice of the appropriate mappings depends mainly on the following word and concept properties:

- part-of-speech tag (given for each word)
- semantic type (given for each concept)

The decision was made that all the words (parts) of a concept, which are not nouns or adjectives are disregarded and not considered in further analysis. Moreover, the selected mapping's concepts must carry one or more semantic types belonging only to the previously defined set of thirty "valid" semantic types (Table 1). There is one conditionally accepted semantic type, namely *Spatial Concept (spoc)*, which is valid only if it is accompanied by other concepts of the valid semantic types. It was verified that in this specific case, especially for the *Body (Brain), Organ or Region* concepts, the *Spatial Concept* semantic type provides a complementary, locational information and cannot be ommited.

These above mentioned conditions are checked in the Mapping Analyzer and Keyword Selector steps. Every concept has one of the following tags assigned:

- N – a concept with all valid semantic types, consisting of at least one noun word.
- J – a concept with all valid semantic types, consisting of only adjective words (no nouns)
- S – a concept with all valid semantic types, consisting of at least one noun word or one adjective word and at least one word with *Spatial Concept (spoc)* semantic type.
- X – all other non-valid concepts (carrying invalid semantic type or part of speech)

Consequently, each mapping, will be a finite chain of the mentioned tags. In the Keyword Selector step tag-chains are verified by a regular expression, which is a pattern for finding valid concepts in the mapping, resulting in creation of final keywords.

Context-dependent query construction

In order to use efficiently any database like PubMed or Google Scholar, a well constructed query must be prepared. In our approach, the keywords from one semantic group are joined together by logical OR operator. At last, all the semantic groups keywords are glued together by AND operator, which gives a final form of a query (Figure 3).

It preserves that at least one of the concepts from each semantic group must be in a retrieved document. As it was mentioned before, a set of semantic groups needs to be defined before searching, what may be understood as a definition of context of interest for a search.

For example, if we are interested in finding similar publications, which mention correspondent brain

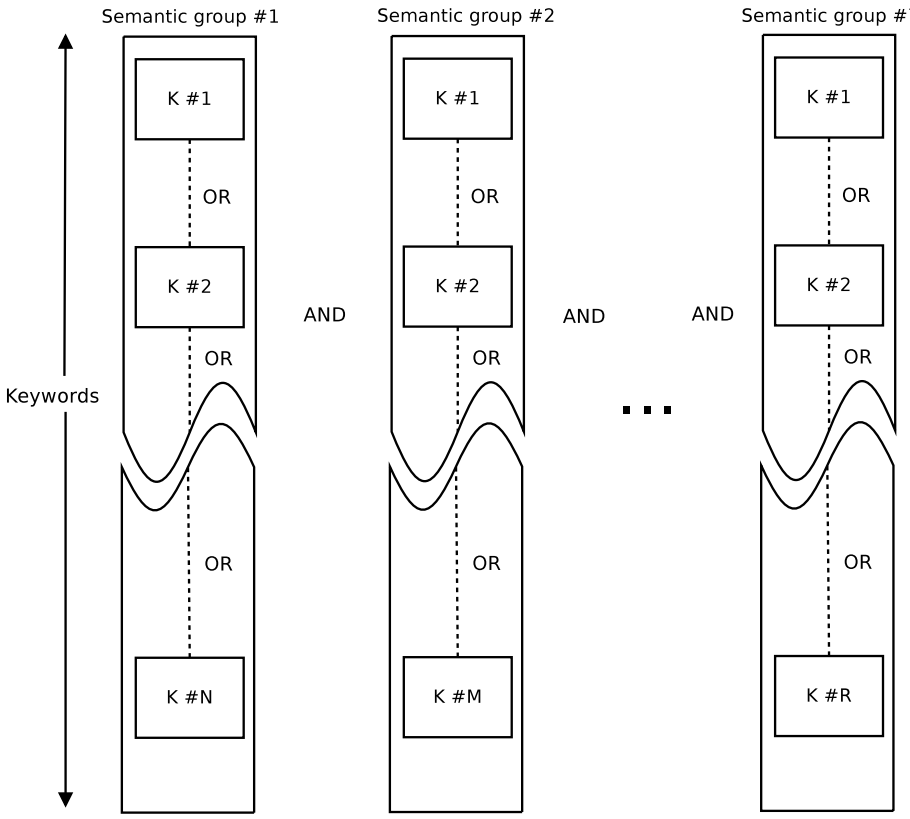


Figure 3: Logical query construction. In the Keywords Extraction step, keywords are classified into one of seven available semantic groups. In the Query Construction step, all the keywords within the same semantic group are joined with the OR logical operator. Then, all the semantic groups are joined with AND logical operator. Later, such query is executed on the PubMed database in order to retrieve relevant publications.

locations and type of procedures employed in the input document, the Brain part or region and Diagnostic procedure semantic groups should be only enabled in a search.

On the other hand, if we want to find publications which refer to similar diseases mentioned in the input document, the Diseases or Syndrome semantic group should be used in a search.

Finally, if there is an intention for finding publications similar in a general way to the input document, the

context of search should be very broad, hence all the semantic groups should be selected in a search.

Query refinement and automatic results analysis

This is a proposal phase, which is considered for testing in the future. After the query is executed and the retrieved publications arrive from PubMed, they may be further analyzed. In the *Results validation* step (Figure 1), the number of retrieved documents is to be ckecked. If a query was too selective (few documents returned) or too broad (thousands of documents returned), a set of procedures should be designed, which re-arrange the original query (*Query tuning* step) to make it less selective or less broad.

This cycle (*Results validation–Query tuning*) should repeat until the moment when either the number of retrieved papers is satisfiable or a pre-defined time threshold has passed.

Flexibility and extensibility of the method

It must be mentioned that even the methodology described here is thought to work in the domain of neuroscience, it may be easily adjusted for work with terminology of any other biomedical domain defined in UMLS. It may be done through appropriate re-definition of "valid" semantic types and semantic groups.

Results and Discussion

In this section we present two examples illustrating the presented methodology:

- Example I. A demonstration of a sample input processing by the methodology. The intermediate results from each of its steps are discussed.
- Example II. A demonstration of the possible integration of the methodology with a coordinate-based database.

The examples base on the randomly selected test paper: "Imitating expressions: emotion-specific neural substrates in facial mimicry." by [20].

The preliminary results discussed by these examples were obtained thanks to the prototype partial implementation of the discussed methodology in Python programming language.

Example I: Input processing

The input (natural language English text) is being processed in few subsequent steps of the methodology shown in Figures 1 and 2.

Let's consider the following sentence (utterance) taken from the abstract of our test paper:

> "Moreover, the magnitude of facial movement during emotion-imitation predicted responses within right insula and motor/premotor cortices."

The analysis of this sentence is discussed below and summarized in Table 2.

Noun phrase	Moreover,		the magnitude		of facial movement		
Tag	Moreover	,	the	magnitude	of	facial	movement
POS	adv	punct	det	noun	prep	adj	noun
Concept				Magnitude		Facial Movement	
Sem. type				qnco		spco	orgf
Keyword						facial movement	

Noun phrase	during emotion-imitation predicted responses					
Tag	during	emotion	-	imitation	predicted	responses
POS	prep	noun	punc	noun	adj	noun
Concept		(*)			Predicted (*)	Emotional Responses
Sem. type		(*)			idcn	(*) menp
Keyword						emotional responses

Noun phrase	within right insula			and
Tag	within	right	insula	and
POS	prep	adj	noun	conj
Concept		Right insula		
Sem. type		bpoc		
Keyword		right insula		

Noun phrase	motor	/premotor cortices.		
Tag	motor	/	premotor	cortices	.
POS	noun	punc	noun	noun	punc
Concept	Motor	Premotor cortex		
Sem. type	ftcn	bpoc		
Keyword		premotor cortex		

Table 2: Processing of the sample sentence (methodology's input). First, Metamap splits the sentence into 8 noun phrases (**'Noun phrase'**). Each noun phrase is further split into tags (**'Tag'**), which later are labeled with appropriate part-of-speech (**'POS'**). Having also the ontological mappings (concepts) and associated semantic types given by Metamap (**'Concept'** and **'Sem. type'**), they are being verified with a regular expression. The sucessfully verified concepts become valid keywords used later in construction of a PubMed query.

In our example, there are four keywords extracted from the analyzed sentence:

facial movement, emotional responses, right insula, premotor cortex

The two last keywords ('right insula' and 'premotor cortex') carry the following UMLS's semantic type: Body Part, Organ, or Organ Component (bpoc). In reference to Table 1 both keywords are classified into the Body (Brain) Part, Organ or Region semantic group.

The 'facial movement' keyword consists of two words: the noun 'movement' carries the Organism Function (orgf) semantic type and the preceding adjective 'facial' carries the Spatial Concept (spoc) semantic type. As a result, this keyword classifies into the Functions semantic group.

Finally, the 'emotional responses' keyword spans two words (adjective and noun) which carry the Mental Process (menp) semantic type. Therefore, it automatically classifies into the Mental Process or Finding semantic group.

There are a few additional concepts mapped by MetaMap, e.g.: 'Predicted', 'Motor' or 'Magnitude', but since none of them carries a "valid" semantic type, they are disregarded.

Next, a logical query is constructed with the four keywords:

((right insula) OR (premotor cortex)) //*Body Part, Organ or Region category*
AND
(facial movement) //*Functions category*
AND
(emotional response) //*Mental Process or Finding category*

According to the previously introduced two proposal steps of the methodology, if a number of publications retrieved with the current query was too small or too big, the query would not be validated and might be modified (tuned). One possible way of query tuning, which we propose, is to join together all single keywords from various semantic groups. In this case the keywords: 'facial movement' (Functions) and 'emotional response' (Mental Process or Finding) are joined together:

((right insula) OR (premotor cortex))
AND
((facial movement) OR (emotional response)) //*joined keywords*

The retrieved set of input-similar documents is the last step of the methodology.

Example II: Integration of the methodology with a coordinate-based database

In this example we test the context-dependent search, provided by the presented approach, on the data retrieved from the coordinate-based database, the Brede database. This database, given a set of brain coordinates (volume), is able to retrieve papers which contain similarly localized volumes [21].

Therefore, the aim here is to employ our methodology to obtain a set of publications, similar, by brain location, to each of four highest ranked documents returned by the Brede database. For the discussion purposes, we additionally perform the same test replacing our methodology with the PubMed retrieval system.

First, from our test paper [20] we got a brain volume represented by a set of coordinates, which, according to the test paper's authors, represent the sites in brain where neural activation was associated with observation of angry faces contrasted with observation of static neutral faces. The original set of coordinates was given in the MNI space, thus we had to transform them to the Talairach space in order to fulfill the Brede Database's requirements. The BredeQuery plugin for SPM was used for this purpose. The transformed coordinates are presented here:

(-42, -5, 29) (12, -5, 42) (-3, 56, -7) (-33, 20, -21)

Later, the Brede database was queried with the coordinates and the papers, reporting the most similar volumes in brain, were retrieved. We present the four highest ranked papers:

1. Karama (2002); "Areas of brain activation in males and females during viewing of erotic film excerpts."
2. Chen (2002); "Spatial summation of pain processing in the human brain as assessed by cerebral event related potentials."
3. Zald (2002); "Brain activity in ventromedial prefrontal cortex correlates with individual differences in negative affect."
4. Pelletier (2003); "Separate neural circuits for primary emotions? Brain activity during self-induced sadness and happiness in professional actors."

Finally, we ran our methodology separately for each of the four papers. Moreover, we tested the methodology with two different settings of semantic groups (see Table 3). In the first setting we selected

four semantic groups (including the Body (Brain) Parts, Organ or Region group) to get a mixed context of search. In the second setting, we selected only the Body (Brain) Parts, Organ or Region semantic group, thus the context of search was limited only to brain locations.

In parallel, we also retrieved sets of related papers (PubMed's Related Articles feature [22]) for each of the same four papers.

Category	Setting #1	Setting #2
Body (Brain) Part, Organ or Region	•	•
Disease or Syndrome	•	○
Functions	•	○
Mental Process or Finding	•	○
Diagnostic Procedure	○	○
Chemical or Substance	○	○
Gene or Genome	○	○

Table 3: Two different search settings of our methodology used in Example II. The filled bullet symbol (•) marks a semantic group which is enabled for searching.

For each set of related papers, either retrieved by our methodology or by PubMed, we checked if the test paper [20] appears in this set. The results are presented in Table 4.

Brede paper		Our methodology	
#	**PubMed Related**	**Setting #1**	**Setting #2**
1.	○	•	•
2.	○	○	○
3.	○	○	•
4.	○	•	•

Table 4: The Example II test's results. The filled bullet (•) means that a test paper was found in the retrieved set of related papers.

Apparently, in case of the search context set to brain locations only (Setting #2), in three of four sets of the retrieved documents the initial test paper was found. This is a promising result in the sense of integration of the coordinate-based databases with our methodology.

In the Setting #1, when the keywords from additional semantic groups affected the search, the results were slightly worse, but still the initial test paper was found in two of four sets of retrieved documents.

PubMed Related did not return the test paper in any of the four searches.

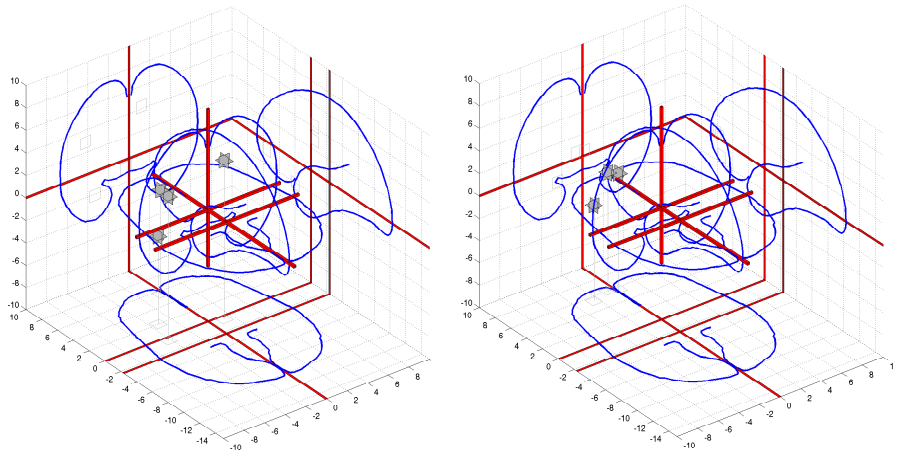


Figure 4: Visualization of two brain coordinate sets (volumes): set of four brain coordinates [(-42, -5, 29) (12, -5, 42) (-3, 56, -7) (-33, 20, -21)] from the test paper (on left) and set of three coordinates [(-24, 54, -14) (-6, 60, 6) (2, 57, 4)] from the Brede paper #4 (on right). According to the Brede database, later confirmed by our methodology, these two coordinate sets (volumes) are very similar. These visualizations confirm that the location of these coordinate volumes is very alike to each other.

Finally, we decided to check how the set of coordinates from the Brede paper #4 differs with the set of coordinates from the test paper. To do that, a visual verification was performed (see Figure 4) using an additional feature of the Brede database service, namely visualization of coordinates in brain. Such visualizations, provided for all the Brede database experiments, are created in Matlab and VRML [23]. It is clearly visible that coordinates (grey stars) in both papers appear in very similar brain locations what is a type of validation for both: the Brede database and our methodology.

Conclusions

We have presented here the general methodology for context-dependent search of related publications, tested in the domain of neuroscience. The input to the system is any English biomedical language text. The whole methodology, represented as a pipeline of subsequent steps, bases on the natural language

processing, logical rules and biomedical ontologies.

Answering the recent need in neuroscience, and more specifically in neuroimaging, for expansion of the results pool in the small coordinate-based databases, we propose the integration of the discussed methodology with the existent tools like the BredeQuery plugin. This plugin, given set of coordinates, employs the Brede database for retrieval of the similar literature directly from the SPM environment. The results presented in Example II confirm the viability of our approach which might be integrated with the above mentioned plugin for further retrieval of related MEDLINE papers.

It should be also mentioned that the recent prototype Python implementation of the methodology might be tested on a coordinate-based database, bigger than the Brede database, e.g. the SumsDB database. In addition, the constructed queries could be directed not only to the PubMed database, but also to other huge comprehensive databases like Google Scholar.

Our future plans, concerning the presented methodology, include implementation of a ranking algorithm which could enable more comprehensive and quantitative evaluation. Two possible solutions for ranking of retrieved papers considered are the Okapi BM25 [24] ranking function and Probabilistic Latent Semantic Indexing [25].

Authors contributions

BW was involved in the design process of the presented methodology and its further prototyping. Moreover, BW carried out a research on the medical ontologies, especially UMLS and related tools. Finally, BW drafted the recent article which was later significantly revised by LKH. LKH introduced the view on the recent neuroimaging needs and was part in the design process of the methodology. Finally, LKH introduced the idea of brain visualizations to this study.

Acknowledgements

This work is supported by Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (CIMBI) – *www.cimbi.org*.

References

1. Van Essen D, Dickson J, Harwell J, Hanlon D: **SumsDB: online access to surface-based representations of cerebral and cerebellar cortex in primates and rodents**. In *Human Brain Project Annual Meeting,*

Bethesda, MD, USA 2004.

2. Laird A, Lancaster J, Fox P: **BrainMap: the social evolution of a human brain mapping database**. *Neuroinformatics* 2005, **3**:65–77.
3. Nielsen F: **The Brede database: a small database for functional neuroimaging**. *NeuroImage* 2003, **19**(2):19–22.
4. Wilkowski B, Szewczyk M, Rasmussen P, Hansen L, Nielsen F: **BredeQuery: coordinate-based meta-analytic search of neuroscientific literature from the SPM environment**. In *Biomedical Engineering Systems and Technologies: International Joint Conference, BIOSTEC 2009, Porto, Portugal, January 14-17, 2009, Revised Selected Papers*, Springer-Verlag New York Inc 2010:314.
5. Crasto C, Marenco L, Migliore M, Mao B, Nadkarni P, Miller P, Shepherd G: **Text mining neuroscience journal articles to populate neuroscience databases**. *Neuroinformatics* 2003, **1**(3):215–237.
6. Crasto C, Masiar P, Miller P: **NeuroExtract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation**. *Journal of the American Medical Informatics Association* 2007, **14**(3):355.
7. Müller H, Rangarajan A, Teal T, Sternberg P: **Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers**. *Neuroinformatics* 2008, **6**(3):195–204.
8. Rindflesch T, Fiszman M: **The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text**. *Journal of Biomedical Informatics* 2003, **36**(6):462–477.
9. McCray A, Srinivasan S, Browne A: **Lexical methods for managing variation in biomedical terminologies.** In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association 1994:235.
10. Smith L, Rindflesch T, Wilbur W: **MedPost: a part-of-speech tagger for biomedical text**. *Bioinformatics* 2004, **20**(14):2320–2321.
11. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch T: **Semantic MEDLINE: a web application to manage the results of PubMed searches**. In *Proceedings of the Third International Symposium for Semantic Mining in Biomedicine* 2008:69–76.
12. Aronson A: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** In *Proceedings of the AMIA Symposium*, American Medical Informatics Association 2001:17–21.
13. Dai M, Shah N, Xuan W, Musen M, Watson S, Athey B, Meng F: **An efficient solution for mapping free text to ontology terms**. *AMIA Summit on Translational Bioinformatics 2008, San Francisco* 2008.
14. Bhatia N, Shah N, Rubin D, Chiang A, Musen M: **Comparing concept recognizers for ontology-based indexing: MGREP vs. MetaMap**. *AMIA Summit on Translational Bioinformatics 2009, San Francisco* 2009.
15. Jonquet C, Shah N, Musen M: **The open biomedical annotator**. *AMIA Summit on Translational Bioinformatics 2009, San Francisco* 2009.
16. Lindberg D, Humphreys B, McCray A: **The Unified Medical Language System**. *Methods of Information in Medicine* 1993, **32**(4):281–291.
17. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic Acids Res* 2004, **32**(Database Issue):D267–270.
18. Humphrey S, Rogers W, Kilicoglu H, Demner-Fushman D, Rindflesch T: **Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment**. *Journal of the American Society for Information Science and Technology* 2006, **57**:96–113.
19. Talairach J, Tournoux P: *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging*. Thieme Medical Publisher Inc., New York 1988.
20. Lee T, Josephs O, Dolan R, Critchley H: **Imitating expressions: emotion-specific neural substrates in facial mimicry**. *Social Cognitive and Affective Neuroscience* 2006, **1**(2):122–135.
21. Nielsen F, Hansen L: **Finding related functional neuroimaging volumes**. *Artificial Intelligence in Medicine* 2004, **30**(2):141–152.

22. Lin J, Wilbur W: **PubMed related articles: a probabilistic topic-based model for content similarity**. *BMC bioinformatics* 2007, **8**:423.

23. Nielsen F, Hansen L: **Experiences with Matlab and VRML in functional neuroimaging visualizations**. In *VDE2000-Visualization Development Environments, Workshop Proceedings, Princeton, New Jersey, USA*, Citeseer 2000:27–28.

24. Robertson S, Zaragoza H, Robertson S, Zaragoza H: **The probabilistic relevance framework: BM25 and beyond**. *Information Retrieval* 2009, **3**(4):333–389.

25. Hofmann T: **Probabilistic latent semantic indexing**. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM 1999:50–57.

Appendix B

BredeQuery: Coordinate-Based Meta-analytic Search of Neuroscientific Literature from the SPM Environment

---

Bartłomiej Wilkowski, Marcin Szewczyk, Peter Mondrup Rasmussen, Lars Kai Hansen, Finn Aarup Nielsen. BredeQuery: Coordinate-Based Meta-analytic Search of Neuroscientific Literature from the SPM Environment. *Part of: Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science: International Joint Conference, BIOSTEC 2009, Porto, Portugal, January 14-17, 2009, Revised Selected Papers*, 314–324, Springer-Verlag New York, 2010. Published.

BredeQuery: coordinate-based meta-analytic search of neuroscientific literature from the SPM environment

Bartłomiej Wilkowski, Marcin Szewczyk, Peter Mondrup Rasmussen, Lars Kai Hansen, and Finn Årup Nielsen

Informatics and Mathematical Modelling, Technical University of Denmark, Kongens Lyngby, Denmark,
bw@imm.dtu.dk,
WWW: http://neuroinf.imm.dtu.dk/

Abstract. Large amounts of neuroimaging studies are collected and have changed our view on human brain function. By integrating multiple studies in meta-analysis a more complete picture is emerging. Brain locations are usually reported as coordinates with reference to a specific brain atlas, thus some of the databases offer so-called coordinate-based searching to the users (e.g. Brede, BrainMap). For such search, the publications, which relate to the brain locations represented by the user coordinates, are retrieved. We present BredeQuery – a plugin for the widely used SPM data analytic pipeline. BredeQuery offers a direct link from SPM to the Brede Database coordinate-based search engine. BredeQuery is able to 'grab' brain location coordinates from the SPM windows and enter them as a query for the Brede Database. Moreover, results of the query can be displayed in a MATLAB window and/or exported directly to some popular bibliographic file formats (BibTeX, Reference Manager, etc).

1 Introduction

The growing number of functional neuroimaging studies of increasingly sophisticated human brain activity brings the demand for new tools/services for integration of research findings, wider exchange of information between laboratories from the same research area and efficient searching of related articles, reviews and other literature [1].

The dominant paradigm in current neuroimaging is that of *functional localization*. Functional localization hypothesizes that a given human behavior is established by a change in brain activity in a relatively limited number of spatially segregated processing units. Thus the result of an experiment under this paradigm consists of a Statistical Parametric Map (SPM) indicating the local involvement. Often the SPM is summarized as a list of regions, see e.g., [2, 3], in which the SPM has been judged to be significantly different from zero (regions were the null hypothesis is rejected). As the typical neuroimaging experiment investigates a highly controlled behavior and often involves a relatively

limited number of subjects, there is strong need for tools to integrate multiple experiments in order to increase the robustness to the experiment specific implementation of the given behavior and to statistical fluctuation due to limited sample sizes.

Several methods have been proposed for neuroimaging meta-analysis and for estimation of associations between the brain locations and textual representations of behavior, for a recent review, see e.g., [1]. A set of methods are based on the so-called Brede Database [4]. Methods for integration include estimation of conditional probability density functions representing the localized probability of activation in response to a given behavior 'word' [5, 6] and multivariate methods based on non-negative matrix factorization that aim to represent global dependencies between brain activation and semantic text labels from neuroscience publications [7].

Brain locations are reported as region coordinates relative to a specific brain atlas (usually MNI or Talairach spaces), hence, there is an interest for effective search for experiments, hence, scientific papers, which report similar coordinate sets in brain. BrainMap [8] and Brede [4] are the databases which offer the coordinate-based searching. For Brede it is available on both the webpage and in a standalone application. A more extensive classification of the databases for fMRI coordinates can be found in [9].

In order to enable a neuroimaging scientists to perform meta-analysis in the context of a specific ongoing study we here propose a tool that integrates retrieval of related research within the data analysis pipeline. The dominant tool for human brain mapping is undisputable the SPM[1] set of tools developed and distributed by the Functional Imaging Laboratory (London, [2]). For an analysis of the usage of imaging pipelines see e.g., [10]. Thus we have initiated the development of a *plugin* for SPM, which offers high integration with the Brede Database.

The *BredeQuery plugin* (see Figure 1) provides the opportunity to perform coordinate-based query and retrieval of the related articles references directly from the SPM (Matlab) environment.

2 Brede Database

The Brede Database available through the webpage[2] records published neuroimaging experiments that list stereotaxic coordinates in so-called MNI or Talairach space [11]. Presently, close to 4000 coordinates from 186 papers with a total of 586 experiments are available.

The data is stored in XML files, and Matlab functions generate static webpages with visualization of the entries in the database, see Figure 2. Web-based searching on coordinates is possible from the homepage, but up till now it has required that the researcher manually typed in the query or extracted results from the image analysis program.

[1] Statistical Parametric Mapping

[2] `http://neuro.imm.dtu.dk/services/brededatabase/`

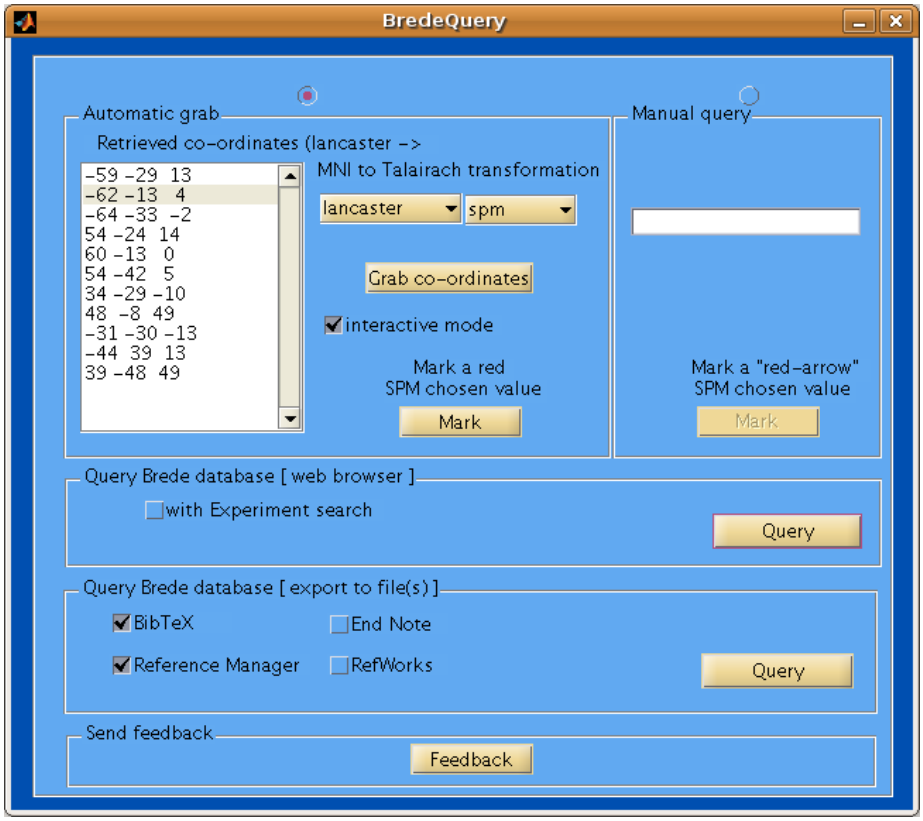


Fig. 1. Graphical user interface of the BredeQuery plugin for SPM. Firstly, the user can choose if the coordinates used for querying will be grabbed from an SPM's results window or will be typed manually. The grabbed (retrieved) coordinates are shown on the list. The user can switch on an interactive mode – the coordinate selected in the SPM window will be automatically selected in the plugin on the coordinates list. Moreover, the coordinates are grabbed using the chosen MNI to Talairach transformation (Brett or Lancaster MTT affine transformations). Afterwards, the user is able to display the query results in the Matlab web browser or to import them into the specified bibliographic format.

The Brede Database web service provides also links to other neuroscientific resources. While querying the database with a specified coordinate in brain, the user is also able to visualize the location in INC Talairach Atlas. Each publication relates by ID number to other databases like PubMed or BrainMap. Brain regions from each of the experiments are mapped to the services like MeSH, BrainInfo, CoCoMac database or Wikipedia. As the Brede webpages are public, the ordinary Web search engines enable text based search of the Brede Database. Furthermore, the researcher may navigate the database via several hyperlinked webpages including brain region, brain function and author ontologies, see Figure 3.

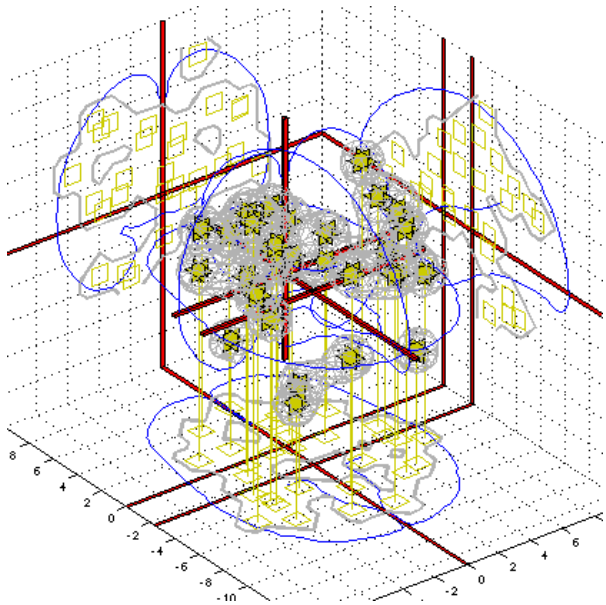


Fig. 2. Screenshot from one of the pages in the Brede Database showing coordinates in Talairach space. This is one of presently 586 experiments recorded in the database – an fMRI experiment resulting in 29 reported coordinates.

3 Related tools

There are a few available tools with similar functionality and aims as the BredeQuery plugin.

The AMAT SPM toolbox was developed by Antonia Hamilton for the Matlab environment. It provides coordinate-based search for over 5000 coordinates from 213 published papers of which some were derived from the Brede Database. The coordinates are in MNI or Talairach space. The toolbox can locate neighboring

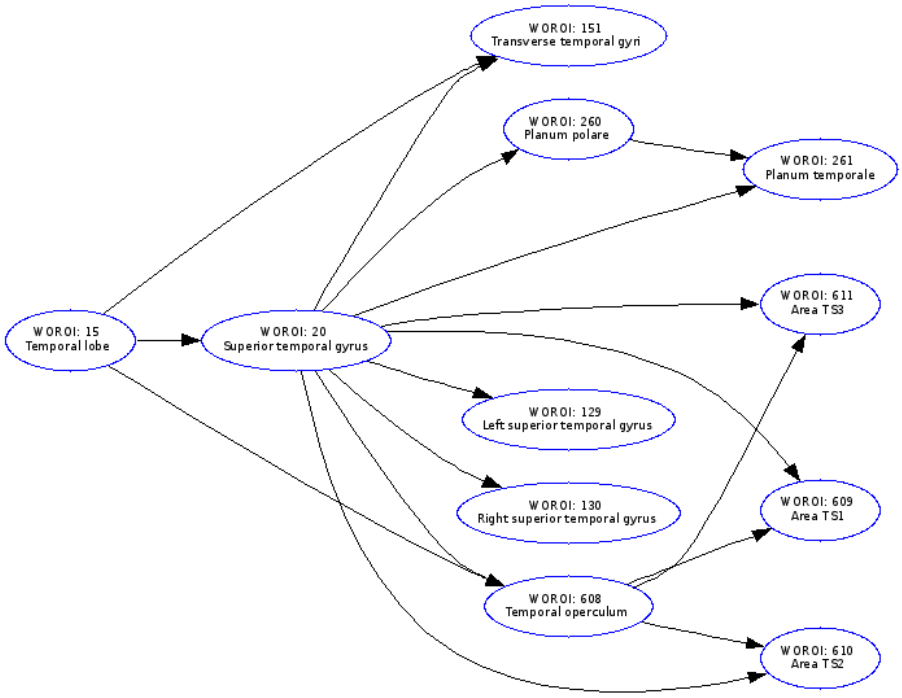


Fig. 3. Relationships and taxonomy of the regions in brain associated with superior temporal gyrus. The entire ontologies for brain regions and brain functions are available together with the Brede Database.

coordinates to a given coordinate, as well as publications for a given author or year. The tool was last updated in 2005 and is available in the internet[3].

Another related toolbox, xjView, offers the SPM user, apart from viewing the images in glass view, section view or 3D render view, search of selected brain regions in databases in order to elucidate their function. It performs the searching among others in Google Scholar[4] and PubMed[5] database. This toolbox was created by Xu Cui and Jian Li and is publicly available[6].

The XCEDE SPM Toolbox [12] enables the users to capture activation data for PET/fMRI analysis and save them to the XML file in a XCEDE XML schema. Moreover, it is extending the exported XML file by automatically adding the anatomical labeling of the region in the brain for the given activity coordinates. It is achieved through two other toolboxes: Talairach Daemon[7] and Automated Anatomical Labeling[8].

4 Software description

The recent version of the *BredeQuery plugin*, together with the User's Guide, can be downloaded from the webpage:

`http://neuroinf.imm.dtu.dk/BredeQuery/`

A graphical user interface of the BredeQuery plugin is divided into five areas where different user-actions can be performed. Firstly, the activation coordinates can be 'grabbed' from the SPM results figure into the plugin. Since the coordinates can be presented in MNI or Talairach spaces, transformations are introduced for interoperability. The coordinate-based search in the Brede Database is based on Talairach space coordinates, thus the BredeQuery plugin offers two MNI to Talairach transformations, which can be chosen by the user. The piecewise affine transformation proposed by Matthew Brett is one of the available transformations [13]. Also included is the affine transformation MNI-to-Talairach (MTT), suggested by Jack Lancaster et al. [14]. Three separate transformations were suggested by his group: one for SPM, one for FSL and a combined 'pooled' transformation. The MTT_{SPM} transformation is set as default in the BredeQuery plugin.

When the coordinates have been 'grabbed' and shown in the BredeQuery plugin, the coordinate-based querying with Brede Database can be done. One or more coordinates can be selected for querying and the results from the Brede Database (publications related to the given activity coordinate) are displayed by the plugin in a web browser (see Figure 4), exported to an XML file or saved

[3] `http://www.antoniahamilton.com/amat.html`
[4] `http://scholar.google.com/`
[5] `http://www.ncbi.nlm.nih.gov/pubmed/`
[6] `http://people.hnl.bcm.tmc.edu/cuixu/xjView/`
[7] `http://www.talairach.org/`
[8] `http://www.cyceron.fr/freeware/`

in the bibliographic file format (BibTeX, Reference Manager, RefWorks or EndNote). We mention that the coordinates need not necessarily be grabbed from SPM in order to make a query. The coordinates can also be entered manually in a manner similar to the functionality of the Brede Database web service.

The user is also able to perform an 'experiment search' (available in the Brede Database service) via the BredeQuery. It has previously been suggested how a similarity can be computed between one set of coordinates and a volume or another set of coordinates [6]. This procedure required the conversion of the set of coordinates to a volume by kernel density estimation. It is, however, not necessary to convert the coordinates to a volume if only the similarity between two coordinates sets are to be compared. It will then generally be faster to compute the similarity based on all coordinate-coordinate pair-wise similarities and perform a weighted summation. There are multiple ways to compute the similarity. Presently, the web-service for the Brede Database uses the following Gaussian/Euclidean form:

$$s_{q,e} = \frac{1}{\sqrt{N}} \sum_{m=1}^{M} \sum_{n=1}^{N} \exp\left(-\frac{(x_{m,q} - x_{n,e})^2 + (y_{m,q} - y_{n,e})^2 + (z_{m,q} - z_{n,e})^2}{2\sigma^2} \right),$$

where σ is set to 10 millimeters, $(x_{m,q}, y_{m,q}, z_{m,q})$ is the mth of M three-dimensional query coordinates, while $(x_{n,e}, y_{n,e}, z_{n,e})$ are the nth of N three-dimensional coordinates in the Brede Database. The factor $1/\sqrt{N}$ aims to regularize for the number of coordinates in each set so that sets with many coordinates do not dominate the search result. A corresponding weight for the query coordinates is not necessary, since this factor will be equal for all queried sets of coordinates of the database.

Following the terminology of BrainMap, a set of coordinates is in the Brede Database called an 'experiment' [15], thus the name 'experiment search'.

The Perl function that presently provides the search functionality to the Brede Database web service is part of the Brede Toolbox, and this toolbox is available on the Internet[9].

5 Example session

In this section we demonstrate the use of the BredeQuery plugin on data from a block-designed auditory fMRI experiment. The experiment was conducted by Geriant Rees, University College London, and the data set was obtained from the SPM webpage[10].

Stimuli were bi-syllabic words, that were presented binaurally. The experimental condition comprised blocks of six scans alternating between rest and

[9] `http://neuro.imm.dtu.dk/software/brede/`

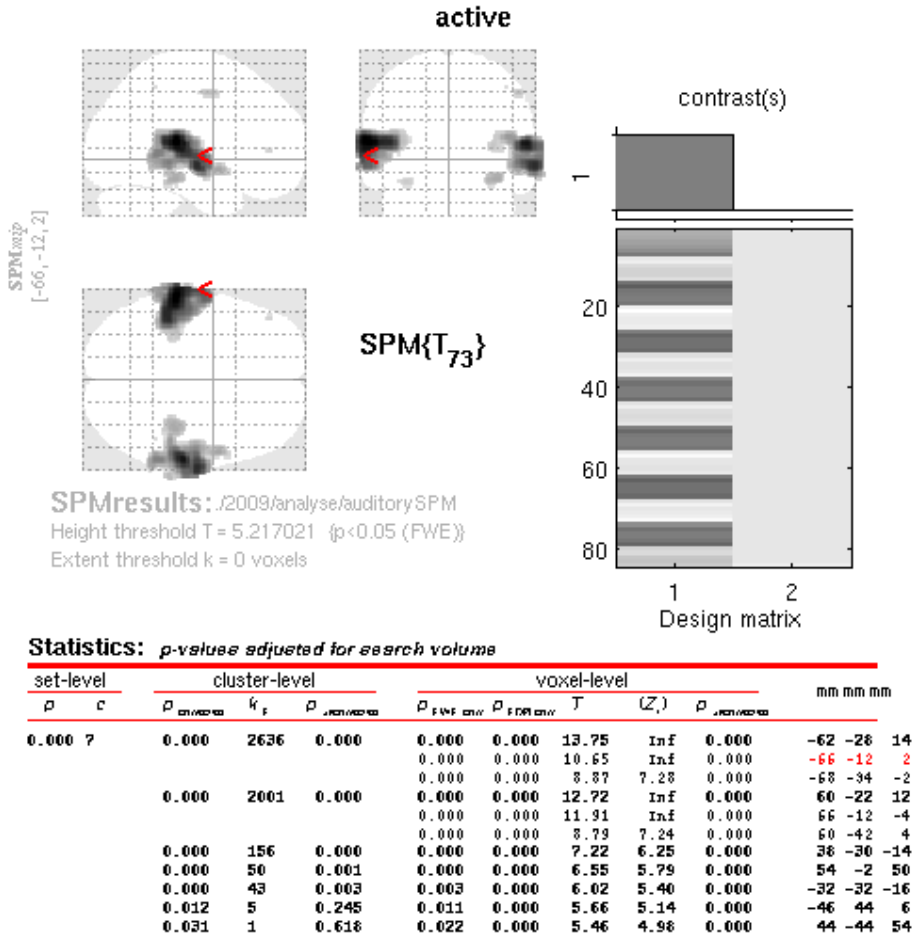
[10] `http://www.fil.ion.ucl.ac.uk/spm/data/auditory/`

#	Distance	x	y	z	WOBIB	Description
1	3.8	-59	-15	2	9	Left medial temporal gyrus — Phonemes (WOEXP: 21)
2	4.1	-62	-14	0	42	Middle superior temporal sulcus — Voice versus non-vocal sounds (WOEXP: 141)
3	4.1	-62	-14	0	42	Middle superior temporal sulcus — Vocal stimuli versus bells, human non-vocal sounds, enveloped white noise and scrampled voices (WOEXP: 142)
4	4.1	-62	-14	0	42	Middle superior temporal sulcus — Frequency-filtered vocal versus non-vocal (WOEXP: 143)
5	5.5	-58	-13	0	88	Left superior temporal gyrus — Activation in sadness film viewing versus neutral film viewing (WOEXP: 282)
6	5.9	-59	-11	-1	123	— Language perception during free viewing (WOEXP: 383)
7	5.9	-61	-13	-1	90	Lateral temporal — Cued recall of familiar people. Group analysis (WOEXP: 290)
8	7.6	-61	-9	-2	64	Left superior temporal gyrus — Listening to voices (WOEXP: 199)
9	7.9	-56	-15	0	88	Left superior temporal gyrus — Activation in sadness film viewing versus amusement film viewing (WOEXP: 284)

Fig. 4. Brede Database query – result displayed in a web browser. List of nearby coordinates to a queried coordinate, displaying distance, the three-dimensional coordinates, the paper identifier, the anatomical label for the retrieved coordinates and short description of the experiment.

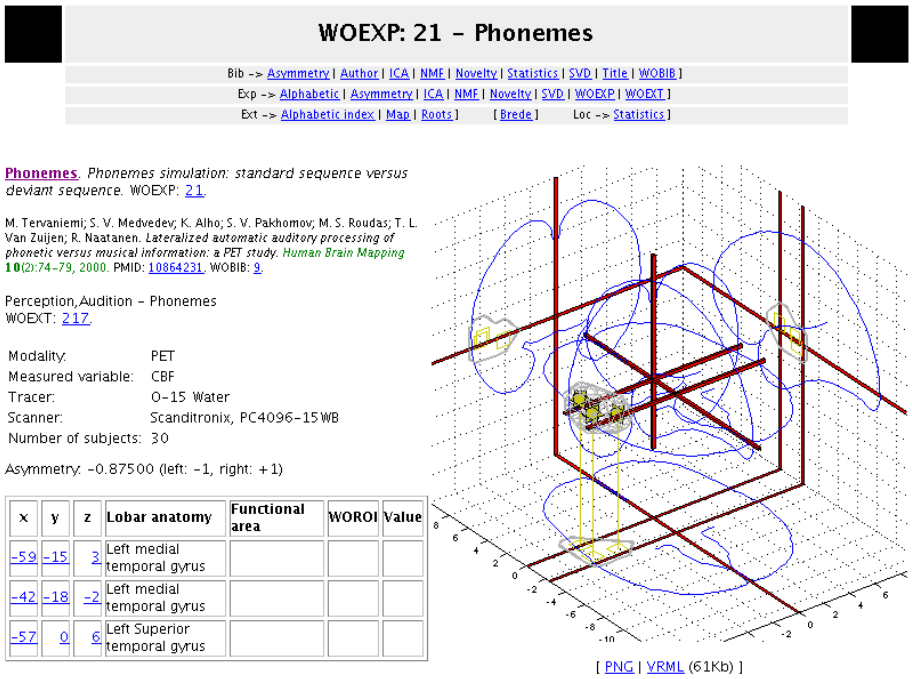
auditory stimulation. Data were preprocessed using the standard SPM pipeline including realignment, spatial normalisation and spatial smoothing. Following preprocessing a conventional univariate statistical analysis was conducted. In the general linear model (GLM) the design matrix comprised a box-car function convolved with the hemodynamic response function (HRF). Figure 5 presents the analysis result. The statistical parametric map was based on a t-contrast (stimulation>rest), with $p < 0.05$ corrected for multiple comparisons using family-wise-error (FWE) correction. A prominent activation was observed in the auditory cortex in the bilateral temporal lobes. An example SPM-BredeQuery user's session leading to abovementioned results may proceed with the following steps:

1. The BredeQuery plugin was loaded by choosing the BredeQuery entry in the SPM's toolbox pop-up menu. All coordinates for significant clusters from the statistical table in the *SPM Graphics windows* were grabbed by the plugin and shown in the coordinates list. They were transformed according to the chosen MNI-to-Talairach transformation. In our example, the coordinates were transformed using the Lancaster's MTT_{SPM} affine transformation. The user was interested in the activation, represented by the coordinate in the MNI space as (-66,-12,2) which was selected in the statistical table – see Figure 5. In the BredeQuery window the user pressed the *Mark red SPM chosen value* button and the previously selected coordinate (-66,-12,2) in MNI space, transformed by the plugin to (-62,-13,4) in Talairach space, was marked in the plugin's coordinates list – see Figure 1.
2. The user has pressed the *Query* button in the *Query Brede database [web browser]* panel (shown on Figure 1) and the webpage with the query results (related articles) has appeared. The user was now able to compare the present results and conclusions with those from the retrieved articles. The webpage results from our example are displayed on Figure 4. Among the first matches from the Brede Database there are coordinates found in the *superior temporal sulcus/gyrus* from experiments with auditory stimulation. A link to the taxonomy of the regions in brain associated with superior temporal gyrus (see Figure 3) was also available. Furthermore, the abstracts of articles related to the experiment were available. Figure 6 represents the detailed, online description of one of the matched experiments.
3. The user wanted to reference some of the articles from the retrieved results in a manuscript. He selected an appropriate bibliographic format, (in this example case 'BibTeX'), pressed *Query* button in the *Query Brede database [export to file(s)]* panel (shown on Figure 1) and the BibTeX file with the references was obtained.
4. The user has discovered a missing feature in the BredeQuery plugin. He thus has pressed the *Feedback* button (Figure 1) and sent a comment to the develop team.



0.000	7	0.000	2636	0.000	0.000	0.000	13.75	Inf	0.000	-62	-28	14
					0.000	0.000	10.65	Inf	0.000	-66	-12	2
					0.000	0.000	8.87	7.28	0.000	-68	-34	-2
		0.000	2001	0.000	0.000	0.000	12.72	Inf	0.000	60	-22	12
					0.000	0.000	11.91	Inf	0.000	66	-12	-4
					0.000	0.000	8.79	7.24	0.000	60	-42	4
		0.000	156	0.000	0.000	0.000	7.22	6.25	0.000	38	-30	-14
		0.000	50	0.001	0.000	0.000	6.55	5.79	0.000	54	-2	50
		0.000	43	0.003	0.003	0.000	6.02	5.40	0.000	-32	-32	-16
		0.012	5	0.245	0.011	0.000	5.66	5.14	0.000	-46	44	6
		0.031	1	0.618	0.022	0.000	5.46	4.98	0.000	44	-44	54

Fig. 5. The demonstration of example results window in SPM. The user can see regions with significant brain activation. The region of activation represented by the coordinate (-66,-12,2) in MNI space is selected. The same coordinate, transformed to the Talairach space using Lancaster's MTT_{SPM} transformation is marked on the BredeQuery's coordinates list as (-62,-13,4) (see Figure 1). Afterwards, the user is able to submit coordinate-based queries to the Brede Database and get the articles related to the same (or nearby) brain regions.



WOEXP: 21 – Phonemes

Bib -> Asymmetry | Author | ICA | NMF | Novelty | Statistics | SVD | Title | WOBIB]
Exp -> Alphabetic | Asymmetry | ICA | NMF | Novelty | SVD | WOEXP | WOEXT]
Ext -> Alphabetic index | Map | Roots] [Brede] Loc -> Statistics]

Phonemes. *Phonemes simulation: standard sequence versus deviant sequence.* WOEXP: 21.

M. Tervaniemi; S. V. Medvedev; K. Alho; S. V. Pakhomov; M. S. Roudas; T. L. Van Zuijen; R. Naatanen. *Lateralized automatic auditory processing of phonetic versus musical information: a PET study. Human Brain Mapping* **10**(2):74-79, 2000. PMID: 10864231. WOBIB: 9.

Perception,Audition – Phonemes
WOEXT: 217.

Modality:	PET
Measured variable:	CBF
Tracer:	O-15 Water
Scanner:	Scanditronix, PC4096-15WB
Number of subjects:	30

Asymmetry: -0.87500 (left: -1, right: +1)

x	y	z	Lobar anatomy	Functional area	WOROI	Value
-59	-15	3	Left medial temporal gyrus			
-42	-18	-2	Left medial temporal gyrus			
-57	0	6	Left Superior temporal gyrus			

[PNG | VRML (61Kb)]

Fig. 6. The online description of an experiment found in the Brede Database. The detailed information about experiment called *Phonemes*, the reference to the article where it was presented, settings of the experiment (modality, type of scanner, tracer used, number of subject, etc.) is displayed. Moreover, all the reported activation coordinates are shown with the indication of the brain location. Finally, the visualization of the coordinates in Talairach space is displayed to the right of the table with coordinates' values.

6 Future work

The first official BredeQuery plugin's version was released to the SPM community on 12th March 2009. Since the plugin is still under development, all incoming feedback comments from the plugin's users are going to be taken into account while releasing updates and further versions, thus more features should be expected.

It was recently emphasized that there are many separated research communities in neuroscience, which do not want to share or exchange the experimental data [16]. Researchers have expressed concerns that sharing of data can lead to unfair use [17]. However, data sharing is important to create trusted collaboration community and is a current topic in debate on future of neuroscience [18, 19] as it is believed that broad data sharing could lead to breaktroughs in our understanding of brain function [20]. Invoking online social networks and computer-based communication can support closer relationships and trust [21] hence, reduce the resistance to data sharing.

Consequently, an interesting extension of the functionality of the plugin can be a direct connection from the SPM environment to a neuroscientific research community, web service or social network. The user would be able to upload the coordinates, results of the analysis, to his own account and save in the assigned server disk space in order to process them later. He can decide whether he wants to keep it private, share only with his research group or alternately release it as a public resource to all users of the service.

It is also possible to employ the BredeQuery plugin to expand the Brede Database. The increment in number of the articles stored in the database could cause bigger interest from the neuroscientists. They could then be encouraged to register their published or unpublished publications in the database via the BredeQuery plugin together with the reported coordinates and keywords.

Finally, we are planning to employ SKEEPMED (Semantic KEyword Extraction Pipeline for MEdical Documents) which is now under development [22]. This pipeline can be used for automatic keyword extraction from abstracts and/or whole papers retrieved by Brede Database. The obtained keywords can be used to query bigger and up-to-date medical databases like PubMed, what consequently could improve the BredeQuery plugin's search results by returning to the user more recent publications related to the respective area in brain and experiments.

7 Conclusions

We have presented herein the BredeQuery plugin for SPM - an application which offers a direct link from the SPM environment to the Brede Database. It provides a mechanism which allows the SPM user to find references to articles which relate to the similar brain activation areas through so-called coordinate-based searching. Moreover, the BredeQuery plugin facilitates the creation of the bibliography files in popular formats.

8 Acknowledgments

We would like to thank Torben Lund and Julian Macoveanu for very constructive comments and feedback. This work is supported by Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (CIMBI) – `www.cimbi.org`.

References

1. Wager, T.D., Lindquist, M., Kaplan, L.: Meta-analysis of functional neuroimaging data: current and future directions. Social Cognitive and Affective Neuroscience **2**(2) (2007) 150–158
2. Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W.: Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press (2007)
3. Pekar, J.: A brief introduction to functional MRI. IEEE Engineering in Medicine and Biology Magazine **25**(2) (March 2006) 24–26
4. Nielsen, F.Å.: The Brede database: a small database for functional neuroimaging. In: NeuroImage. Volume 19., Elsevier (June 2003) Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19-22, 2003, New York, NY.
5. Nielsen, F.Å., Hansen, L.K.: Modeling of activation data in the BrainMap(TM) database: Detection of outliers. Human Brain Mapping **15**(3) (March 2002) 146–156
6. Nielsen, F.Å., Hansen, L.K.: Finding related functional neuroimaging volumes. Artificial Intelligence in Medicine **30**(2) (2004) 141–151
7. Nielsen, F.Å., Hansen, L.K., Balslev, D.: Mining for associations between text and brain activation in a functional neuroimaging database. Neuroinformatics **2**(4) (dec 2004) 369–380
8. Laird, A.R., Lancaster, J.L., Fox, P.T.: BrainMap: The Social Evolution of a Human Brain Mapping Database. Neuroinformatics **3**(1) (2005) 65–78
9. Derrfuss, J., Mar, R.: Lost in localization: The need for a universal coordinate database. NeuroImage, doi **10** (2009)
10. Nielsen, F.Å., Christensen, M.S., Madsen, K.M., Lund, T.E., Hansen, L.K.: fMRI Neuroinformatics. IEEE Engineering in Medicine and Biology Magazine **25**(2) (March 2006) 112–119
11. Talairach, J., Tournoux, P.: Co-planar Stereotaxic Atlas of the Human Brain. Thieme Medical Publisher Inc, New York (January 1988)
12. Keator, D.B., Gadde, S., Grethe, J.S., Taylor, D.V., Potkin, S.G.a.: A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels. Neuroinformatics **4**(2) (2006) 199–212
13. Brett, M.: The MNI brain and the Talairach atlas. MRC Cognition and Brain Sciences Unit (1999)
14. Lancaster, J.L., Tordesillas-Gutiérrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., Mazziotta, J.C., Fox, P.T.: Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. Human Brain Mapping (January 2007)
15. Fox, P.T., Mikiten, S., Davis, G., Lancaster, J.L.: BrainMap: A database of human function brain mapping. In Thatcher, R.W., Hallett, M., Zeffiro, T., John, E.R., Huerta, M., eds.: Functional Neuroimaging: Technical Foundations. Academic Press, San Diego, California (1994) 95–105

16. Ascoli, G.A.: The Ups and Downs of Neuroscience Shares. Neuroinformatics **4** (September 2006) 213–216(4)
17. Teeters, J.L., Harris, K.D., Millman, K.J., Olshausen, B.A., Sommer, F.T.: Data Sharing for Computational Neuroscience. Neuroinformatics (February 2008)
18. Kennedy, D.N.: Neuroinformatics and the Society for Neuroscience. Neuroinformatics **5** (September 2007) 141–142
19. Liu, Y., Ascoli, G.A.: Value Added by Data Sharing: Long-Term Potentiation of Neuroscience Research: A Commentary on the 2007 SfN Satellite Symposium on Data Sharing. Neuroinformatics **5** (September 2007) 143–145(3)
20. Van Horn, J.D., Ball, C.A.: Domain-Specific Data Sharing in Neuroscience: What Do We Have to Learn from Each Other? Neuroinformatics **6**(2) (2008) 117–121
21. Lampe, C., Ellison, N., Steinfield, C.: A face(book) in the crowd: social searching vs. social browsing. In: CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, New York, NY, USA, ACM Press (2006) 167–170
22. Wilkowski, B., Szewczyk, M.M., Hansen, L.K.: Bridging the gap between coordinate- and keyword- based search of neuroscientific databases by UMLS-assisted semantic keyword extraction (2009)

Appendix C

Graph-Based Methods for Discovery Browsing with Semantic Predications

Bartłomiej Wilkowski, Marcelo Fiszman, Christopher M. Miller, Dimitar Hristovski, Sivaram Arabandi, Graciela Rosemblat, Thomas C. Rindflesch. Graph-Based Methods for Discovery Browsing with Semantic Predications. *American Medical Informatics Association Annual Symposium*, 1514–1523, Washington D.C., 2011. Published.

Graph-Based Methods for Discovery Browsing with Semantic Predications

Bartłomiej Wilkowski, M.S.,[1] Marcelo Fiszman, M.D., Ph.D.,[2]
Christopher M. Miller, M.D.,[2] Dimitar Hristovski, Ph.D.,[3] Sivaram Arabandi, M.D., M.S.,[2]
Graciela Rosemblat, Ph.D.,[2] Thomas C. Rindflesch, Ph.D.[2]
[1]Technical University of Denmark, DTU Informatics, Kongens Lyngby, DK-2800, Denmark;
[2]National Library of Medicine, Bethesda, MD 20894, USA; [3]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, Ljubljana, Slovenia

Abstract

We present an extension to literature-based discovery that goes beyond making discoveries to a principled way of navigating through selected aspects of some biomedical domain. The method is a type of "discovery browsing" that guides the user through the research literature on a specified phenomenon. Poorly understood relationships may be explored through novel points of view, and potentially interesting relationships need not be known ahead of time. In a process of "cooperative reciprocity" the user iteratively focuses system output, thus controlling the large number of relationships often generated in literature-based discovery systems. The underlying technology exploits SemRep semantic predications represented as a graph of interconnected nodes (predication arguments) and edges (predicates). The system suggests paths in this graph, which represent chains of relationships. The methodology is illustrated with depressive disorder and focuses on the interaction of inflammation, circadian phenomena, and the neurotransmitter norepinephrine. Insight provided may contribute to enhanced understanding of the pathophysiology, treatment, and prevention of this disorder.

Introduction

Sophisticated methods are needed to supplement traditional information retrieval tools for effectively exploiting the large amount of online textual resources currently available. An active area of research in biomedicine in this regard is literature-based discovery (LBD), the primary goal of which is to help researchers make new discoveries by generating novel hypotheses. As pioneered by Swanson,[1] the basic underlying principle of the LBD paradigm is that relations $A - B$ and $B - C$ may be known, yet relation $A - C$ has gone unnoticed. Earlier LBD systems[2, 3, 4] used concept cooccurrence as their primary mechanism for representing relations. Since only some cooccurrences underlie "interesting" relations, this has drawbacks, which have been addressed first by Hristovski et al.[5] and later by Cohen et al.[6] with the use of semantic relations. The use of discovery patterns[5] is a further refinement for focusing on useful relations. One such pattern[5] is *Maybe_treats*, which says (in part) that a therapeutic agent C maybe treats disease A if the level of an important measurement B is typically increased in patients with disease A and if C is able to reduce the level of B. Additional discovery patterns have been investigated.[7, 8]

We present a novel LBD methodology incorporating semantic predications and graph-based methods in order to guide researchers through the relevant literature on a user-specified biomedical phenomenon. The motivation is to extend LDB methodology beyond making discoveries to a principled way of navigating through selected aspects of some research area.[9] An additional goal is to go beyond document retrieval in response to a query by revealing crucial relationships in the domain, which may evolve as the user exploits the method. Related work in network analysis of microarray data is becoming widely used in systems-based research for drug discovery.[10]

LBD provides the ability to uncover previously implicit or unnoticed relationships in the research literature, but has been primarily used when component relationships of the final discovery are already known. In the method we propose, it is not necessary to know ahead of time which relationships may be useful for guiding the research process. Only the general content area need be specified. The method might be thought of as "discovery browsing," using graph theoretic paths to generalize Hristovski's discovery patterns, not with the purpose of necessarily making a discovery, but of explicating poorly understood relationships, by providing novel points of view on some research problem. The method involves an interaction between user decisions and system results. Such "cooperative reciprocity" focuses system output iteratively based on stipulations that bring relevant relations into clearer focus by narrowing choices, thus controlling the explosion of potential relationships often generated in LBD.

The underlying technology depends on semantic predications extracted from MEDLINE citations using SemRep[11] and represented as a large graph of interconnected nodes (predication arguments) and edges (predicates). The graph-theoretic constructs degree centrality and path analysis are used to suggest paths in this graph, which represent chains of relationships that may guide the research process. The methodology is illustrated with selected aspects of depressive disorder.

Background

Semantic predications

The predications for this study were provided by the SemRep system,[11] which relies on biomedical domain knowledge in the Unified Medical Language System (UMLS). Access to the UMLS Metathesaurus is provided by MetaMap,[12] while a set of semantic relationships (predications) is extracted based on the UMLS Semantic Network. For example, SemRep extracts the relationship: "Leptin STIMULATES Serotonin" from text "CNS serotonin activated by leptin modulates sympathetic outflow to the skeleton." Although coverage of text processed by SemRep is a limiting factor in this work, the expressiveness of semantic predications positively contributes to the discovery process. Several evaluations of SemRep (e.g. Ahlers et al.[13]) indicate that precision is near 75%. SemRep has been used to extract nearly 25 million semantic predication instances from some 7 million MEDLINE citations (titles and abstracts dating from 1999 through 2010). These are stored in a MySQL database, which was exploited for this work.

Literature-based discovery

Swanson's[1] paradigm is based on concepts A and C coming from two different, nonoverlapping, domains. The goal is to find an intermediate concept B, which occurs with both A and C, and validates a new, earlier unknown, $A - C$ relationship. Such a discovery is called an open discovery. Another type of discovery, a closed discovery, assumes that a relationship $A - C$ is known. Then, a common concept B and relations $A - B$ and $B - C$ are to be found in order to explicate the relationship $A - C$.

The aim of our work is to expand the B element of the $A - B - C$ paradigm. Our methodology considers B not as a single concept, but as a subchain of intermediate concepts, where $A - B - C$ has the form in (1), where $n \in [1, \infty)$.

$$A - (B_1 - B_2 - B_3 - ... - B_n) - C \tag{1}$$

Investigation of possible chains of relationships may result in a discovery (either open or closed). Swanson[1] introduced the possibility of having a chain of Bs between A and C, but this has not been extensively exploited. We use semantic predications to represent these relationships, and we decided to use graphs as a medium for representing the predications. Referring to graph theory terminology, we implement discovery chains as paths in a graph.

Graph theory

A graph is a representation of connections (edges) between objects (nodes). Graphs, also known as networks, are extensively studied in social network analysis and the Semantic Web. Graph theory is a set of functions and measures pertaining to graph properties. One such measure used in this paper is degree centrality, which measures the connectedness of nodes in a graph. A node with more connections (relationships) to other nodes has higher degree centrality. Freeman[14] describes degree centrality as an indicator of the communication activity in a social network. The importance of high connectivity is also seen in gene interaction networks (p53, for example).[15] In our case degree centrality may be considered as an indicator of the principal substances in the domain for which the graph was constructed. The formula for degree centrality of node v in a graph with n nodes is:[14]

$$C_d(v) = \frac{deg(v)}{n-1} \tag{2}$$

In graph theory, a path is a sequence of edges connecting any two nodes in the graph. Paths may be of any length. The

shortest is of length 1:

$$A - B \tag{3}$$

The longest is of length $N - 1$, where N is the number of nodes in the graph:

$$X_1 - X_2 - ... - X_N \tag{4}$$

In Semantic Web research on ranking paths of semantic associations Anyanwu et al.[16] exploit the notion of "predictability." In their results longer paths more likely reveal rare and uncommon associations.

Dupont et al.[17] discuss many walking approaches in a graph (edge passages), which may be also understood as extraction of paths from the graph. The definitions of maximal length of the edge passage (k-walk) and nodes of interest are based on this work. The nodes of interest are the start and end points of a walk in a graph. For them, length of the walk is the number of intermediate nodes visited during a walk between nodes of interest. We measure path length by the number of edges between the start and end nodes.

Methods

Overview

The procedure for exploiting paths in a graph to facilitate discovery browsing involves several steps: creating a graph of relevant predications, extracting and ranking paths, and finally, inspecting a small subgraph based on selected paths. At several steps in the process, system output is filtered based on user stipulation, representing the cooperative reciprocity involved in uncovering research insights in the domain. A crucial assumption of the system is that the user brings to bear domain knowledge as part of the process of navigating and focusing in the selected area of interest.

Creating the initial graph is an iterative process in which the user specifies a seed concept to extract predications from the SemRep predication database. (For this project, extracted predications were limited to those with one of the substance interaction predicates: STIMULATES, INHIBITS, INTERACTS_WITH, and COEXISTS_WITH.) Concepts in the graph are ranked by degree centrality, and a new seed concept is selected from those highest on the list, which is used to extract additional predications to be added to the growing graph. When a graph of sufficient size to produce "interesting" results has been generated, paths between stipulated concepts are extracted and ranked, also based on degree centrality. Finally, the user selects paths for further analysis. We will illustrate system processing with depressive disorder as the domain of interest.

Create graph

Serotonin was selected as the seed concept for investigating depression, since it is known to be a prominent neurotransmitter in this disorder. We extracted all predications from the database that had an argument containing the string "seroton" (ignoring case). In addition to "Serotonin" this included 183 concepts, such as "serotonin receptor," "Serotonin Agonists," "Serotonin Agents," and "Serotonin 5-HT-3 Receptor." (This was an implementation expedient. In the future, ontology resources will be exploited.) Retrieved predications were loaded into a graph consisting of 1561 nodes (concepts) and 7061 edges (predications) using the NetworkX[18] software package written in the Python programming language. A path of length one in this graph represents all the predication instances in the database having the two nodes as either subject or object. For example, the path "Estradiol-stimulates-Serotonin" represents two instances of the corresponding predication extracted from MEDLINE citations (PMIDs: 16736471 and 19168037). Links between an edge in the graph and sentences in citations from which corresponding predications were extracted are maintained by the system.

As a first step in expanding this graph, we calculated degree centrality and ranked the results. Ignoring the concepts containing "serotonin," melatonin was high on this list, and was chosen to expand the serotonin graph. This was a user choice which focused the subsequent graph on a particular aspect of depression. Other substances high on the degree centrality list, which could have been selected, were "Estrogens," "Dopamine," and "Corticosterone." We then

retrieved all predications from the SemRep predication database that had an argument containing the string "melato" (ignoring case). In addition to "Melatonin," this included 126 concepts, such as "Melatonin Receptors" and "Receptor, Melatonin, MT2." The resultant graph (containing both serotonin- and melatonin-related concepts) consisted of 2207 nodes (concepts) and 11,752 edges (predications).

The growing graph of predications for depression was expanded a third (and final) time. Degree centrality was again calculated and the results ranked. Before prominent concepts were selected from this list, a stop list (based on user domain knowledge) was applied to remove uninformative concepts (e.g. "Pharmaceutical Preparations"), classes of both drugs and body substances (e.g. "Antidepressive Agents," "agonists"), and physiologically general terms (e.g. "Water," "Oils"). Serotonin- and melatonin-related concepts were also ignored since predications with them were already in the graph. The top 140 concepts from this filtered degree centrality list were used in another query to the predication database and the retrieved predications were added to the graph. The resulting graph of predications for the depression study, consisting of 22,828 nodes (concepts) and 435,437 edges (predications), formed the basis for further processing.

Extract paths

The next step was to extract paths from the graph, which, as in constructing the graph, involved an interaction of system output and user stipulations. Serotonin and melatonin were selected as anchors, and all paths of length four between them were extracted from the graph using the depth-first algorithm.[19] This value was selected as a compromise. Longer paths are likely to provide more revealing results for discovery;[16] however, considerations of processing time with the current implementation imposed this limitation. The total number of paths extracted was 4,206,647, and all had the form:

$$[Melatonin(A)] - B_1 - B_2 - B_3 - [Serotonin(C)] \tag{5}$$

Before proceeding, we removed all paths in which one of the five concepts was on the stop list of general and uninformative concepts noted above. (Such concepts "crept" into the graph by being arguments of predications with non-stopped concepts.) 3,840,958 remained, and a composite degree centrality score was computed for each. This was calculated as the arithmetic sum of the degree centrality values for all five nodes in the path:

$$score = \sum_n dc(B_n) \tag{6}$$

The list of paths was ranked based on the composite degree centrality score.

As a further step in limiting the number of paths for analysis, we selected only those containing the concept "CLOCK," based on domain knowledge that recent research implicates the clock genes in depression.[20] The remaining paths (16,141) had one of the following patterns:

$$[Melatonin(A)] - [CLOCK(B_1)] - B_2 - B_3 - [Serotonin(C)] \tag{7}$$

$$[Melatonin(A)] - B_1 - [CLOCK(B_2)] - B_3 - [Serotonin(C)] \tag{8}$$

$$[Melatonin(A)] - B_1 - B_2 - [CLOCK(B_3)] - [Serotonin(C)] \tag{9}$$

We further focused this list by eliminating concepts that we chose not to consider at this point. Paths with concepts such as "Glucose," "Antibodies," "Lipopolysaccharides," "Kinases" etc. were removed (again, based on user domain knowledge), leaving 5,406 paths.

Analyze paths

From this list we chose the top twenty paths (sorted by composite degree centrality) for further analysis.

1. [Melatonin]-[Interleukin-1 beta]-[Glutamate]-[CLOCK]-[Serotonin]
2. [Melatonin]-[CLOCK]-[Glutamate]-[Interleukin-1 beta]-[Serotonin]
3. [Melatonin]-[Insulin]-[Glutamate]-[CLOCK]-[Serotonin]
4. [Melatonin]-[CLOCK]-[Glutamate]-[Insulin]-[Serotonin]
5. [Melatonin]-[Interleukin-6]-[Glutamate]-[CLOCK]-[Serotonin]
6. [Melatonin]-[CLOCK]-[Glutamate]-[Interleukin-6]-[Serotonin]
7. [Melatonin]-[Interleukin-1 beta]-[Norepinephrine]-[CLOCK]-[Serotonin]
8. [Melatonin]-[CLOCK]-[Norepinephrine]-[Interleukin-1 beta]-[Serotonin]
9. [Melatonin]-[Cholesterol]-[Glutamate]-[CLOCK]-[Serotonin]
10. [Melatonin]-[CLOCK]-[Glutamate]-[Cholesterol]-[Serotonin]
11. [Melatonin]-[Insulin]-[Norepinephrine]-[CLOCK]-[Serotonin]
12. [Melatonin]-[CLOCK]-[Norepinephrine]-[Insulin]-[Serotonin]
13. [Melatonin]-[Interleukin-1 beta]-[Interferon Type II]-[CLOCK]-[Serotonin]
14. [Melatonin]-[Interleukin-6]-[Norepinephrine]-[CLOCK]-[Serotonin]
15. [Melatonin]-[CLOCK]-[Norepinephrine]-[Interleukin-6]-[Serotonin]
16. [Melatonin]-[Insulin]-[Interferon Type II]-[CLOCK]-[Serotonin]
17. [Melatonin]-[CLOCK]-[Dopamine]-[Interleukin-1 beta]-[Serotonin]
18. [Melatonin]-[Interleukin-1 beta]-[Dopamine]-[CLOCK]-[Serotonin]
19. [Melatonin]-[Interleukin-6]-[Interferon Type II]-[CLOCK]-[Serotonin]
20. [Melatonin]-[Insulin]-[Dopamine]-[CLOCK]-[Serotonin]

Eleven unique concepts are involved in these paths: “Interleukin-1 beta,” “Interleukin-6,” “Glutamate,” “Dopamine,” “Norepinephrine,” “Insulin,” “Cholesterol,” and “Interferon type II” (in addition to the stipulated “Serotonin,” “Melatonin,” and “CLOCK”). The graphical representation of these paths is shown in Figure 1. Numbers on the edges in Figure 1 show the number of predication instances (limited to substance interaction predicates) represented by that edge. Links to the corresponding citation sentences are maintained by the system and underpin the guidance provided to the user. It should be noted that there is considerable overlap among the predications represented in the paths. Edges with fewer than ten predications are dashed in this graph. We then filtered the subgraph further to include only those connections that occurred ten times or more, thus eliminating “Dopamine.” This filtering also highlights the more direct connection between CLOCK and melatonin, which was not the initial intent of this exploration, but nevertheless is an indication that this visualization highlights important known relationships.

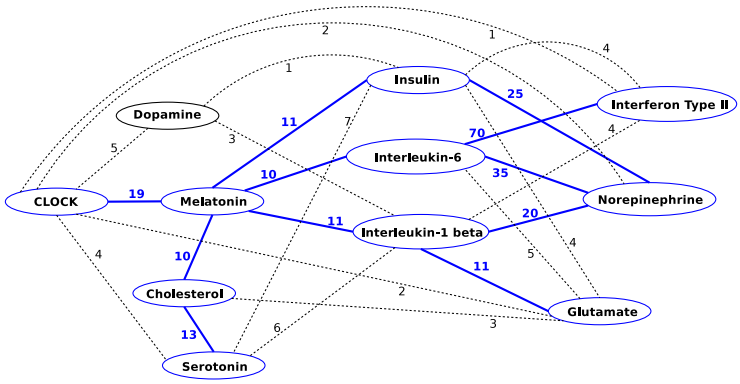


Figure 1: Graph representing the top twenty paths. Numbers on the edges show the number of predication instances.

There are several "stories" implicit in Figure 1. We further decided to concentrate on just one, namely the interactions among melatonin, the clock genes, interleukin-1 beta (IL-1 beta), interleukin-6 (IL-6), and norepinephrine, as shown in Figure 2.

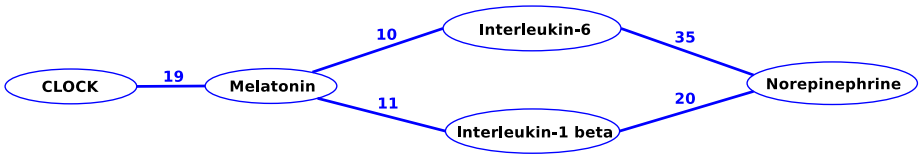


Figure 2: User selected subgraph.

Results

Each of the 95 predications found in Figure 2 was analyzed for SemRep accuracy, and it was determined that 88% were correct. Each of the citations from which these predications had been extracted was then inspected. In general we found that even incorrect predications often reveal relevant, perhaps unknown, relationships that can in turn lead to further research. Below we give examples of predications drawn from citations that provide insight into the three aspects of depression covered by Figure 2: melatonin and CLOCK, melatonin and the two proinflammatory cytokines, the two cytokines and norepinephrine.

CLOCK stimulates Melatonin: Although this predication is wrong, the citation from which it was extracted provides important information about the relationship. "This interaction, . . . may reflect the central role of melatonin, i.e. in synchronising peripheral clock cells that require unique phasing of output signals with the master clock in the brain."[21]

Melatonin interacts with CLOCK / CLOCK interacts with Melatonin: ". . . whereas an 'internal coincidence model' best explains the way melatonin affects the phasing of clock gene expression . . . "[22] "In mammals, the nocturnal rise in pineal melatonin is regulated by signals from the endogenous clock . . . "[23]

Melatonin inhibits Interleukin-1 beta / Melatonin inhibits Interleukin-6 / Melatonin stimulates Interleukin-1 beta / Melatonin stimulates Interleukin-6: "Further melatonin repressed the upregulated levels of expression of proinflammatory cytokines like, TNF-alpha, IL-1beta and IL-6 in RE." (in experimental reflux esophagitis)[24] "Treatment with melatonin significantly increased the levels of IL- 1beta, IL-6, . . . " (in collagen-induced arthritis)[25]

Several predications are about the effect of norepinephrine on either interleukin-1 beta or interleukin-6. We did not pursue this relationship in this project.

Interleukin-6 stimulates Norepinephrine: For the effect of IL-6 on norepinephrine, all of the predications were incorrect. Some of them are nonetheless useful in reporting on a reciprocal relationship between IL-6 and norepinephrine, even if a direct interaction is not noted. "In addition, plasma levels of IL-6 and IL-2 were increased in four stress groups, serum norepinephrine and dopamine were decreased dramatically in stress group and stress low-dose GTPs modulation group."[26]

Interleukin-1 beta inhibits Norepinephrine: "These results suggest that IL-1 beta could decrease NE levels"[27] ". . . IL-1beta-induced suppression of the LH surge is most probably mediated through an increase in GABA levels in the MPA which causes a reduction in NE levels."[28]

Interleukin-1 beta stimulates Norepinephrine: "While acute treatment with IL-1beta increased NE concentrations in both the paraventricular nucleus and the median eminence (ME), chronic treatment increased NE concentrations only in the ME."[29] "These results indicate that IL-1beta increases NE levels both in the PVN and in the ME and this could be a possible mechanism by which it stimulates the HPA axis."[30] "We observed that IL-1beta increases the release of NPY, norepinephrine (NE), and epinephrine (EP) from human chromaffin cells."[31]

Discussion

The methodology presented was illustrated with selected aspects of depressive disorder and focuses on the interaction of inflammation (particularly the proinflammatory cytokines interleukin-1 beta and interleukin-6), circadian phenomena (clock genes and melatonin), and the neurotransmitter norepinephrine. Below we give an overview of the extent of current research on these aspects of depression (PubMed queries issued on 03/10/2011) and suggest how our results may contribute to an understanding of the pathophysiology of this disorder.

There is considerable research investigating circadian phenomena and depression. The PubMed query "(circadian rhythms[mh] OR clock OR melatonin) AND depression[mh]" returns 331 citations. For example, Kennaway[20] reports on the clock genes and behavioral disorders, including depression. Melatonin figures prominently in this review, but a mechanism is not proposed. Rosenwasser[32] reviews research on the clock genes and psychiatric disorders more generally, but mechanisms and the connection with inflammation are not highlighted. Our results provide considerable detail on the mechanisms involved.

Far less research has investigated the interaction of circadian phenomena and inflammation with respect to depression. The PubMed query "cytokine[mh] AND (circadian rhythms[mh] OR clock OR melatonin) AND depression[mh]" only returns 18 citations. When limited to reviews, only 8 are retrieved with this query; several are concerned with specific disorders, such as rheumatologic disorders[33] and cancer.[34] Only one covers general considerations of the interaction of circadian phenomena and inflammation.[35] Our results suggest details concerning the interaction of melatonin and the two proinflammatory cytokines interleukin-1 beta and interleukin-6.

The role of inflammation in depression has been extensively studied. The PubMed query "(inflammation OR cytokines) AND depression[mh]" returns 1500 citations (41 reviews). Raison et al.[36] and Anisman,[37] for example, provide particularly lucid overviews. Our results are not to be thought of as uncovering the insight that inflammation is intimately connected with depression (for which there is considerable evidence), but rather they provide additional information about the interaction of specific cytokines (IL-1 beta and IL-6) and norepinephrine.

Although the noradrenergic system is known to be involved in depression, the mechanistic details are still being investigated.[38] Norepinephrine has not been targeted as intensely as serotonin in therapeutic approaches.[39, 40] Only one norepinephrine reuptake inhibitor (duloxetine) is currently prescribed (in Europe, not in the U.S.).[39] Recently, combined serotonin and norepinephrine reuptake inhibitors are being used.[41, 42] Although it is known that the cytokines interact with norepinephrine,[43] the particular mechanism of IL-1 beta and IL-6 has not been intensively studied (PubMed query "norepinephrine AND (interleukin-1 beta OR IL-1 beta OR interluekin-6 OR IL-6) AND depression" returns 35 citations). Our results point to considerable evidence of the interaction of IL-6 and IL-1 beta (in particular) in a variety of contexts beyond cerebral structures, and thus may suggest new avenues for research in explicating the details.

Finally, there is very little research investigating the comprehensive interaction of inflammatory processes, circadian phenomena, and noradrenergic aspects of depression. The specific PubMed query "(interleukin-1 beta OR IL-1 beta OR interluekin-6 OR IL-6) AND melatonin AND norepinephrine AND depression" returns no citations. The more general "cytokines AND (circadian rhythms[mh] OR clock OR melatonin) AND norepinephrine AND depression[mh]" returns three citations. One of these[39] is a clinically oriented review and does not discuss any of the mechanisms addressed in this paper. The other two[44, 45] cover research more generally, and report substance levels consistent with our results, but do not suggest mechanisms.

We approached high connectedness by filtering to include only those connections with 10 or more occurrences. One weakness of this strategy is that it focuses on the more often studied relations at the expense of others. An alternative approach would be to look at this from the opposite angle, and calculate degree centrality using relations with fewer occurrences each. This may highlight areas that are highly connected but have not been explored in a coordinated manner. A second aspect that can be explored is by taking into account the directionality of the relations – incoming and outgoing relations – and calculating directed paths.

Conclusion

We introduced a novel LBD methodology incorporating semantic predications and graph analysis that guides researchers through the research literature on a user-specified biomedical phenomenon. A type of "discovery browsing" exploits graph theoretic paths in order to elucidate poorly understood relationships by providing novel points of view on some research problem. The user is not required to specify which relationships may serve as a useful guide. A core aspect of the method is that the user brings to bear domain knowledge as part of the process of navigating in the selected area of interest. Such "cooperative reciprocity" focuses system output iteratively, thus controlling the explosion of potential relationships often generated in LBD.

In the method, relationships in MEDLINE citations are represented as a (large) graph of interconnected semantic predications. The system suggests paths in this graph, which represent interesting chains of relationships. The underlying technology depends on semantic predications provided by SemRep, and the graph theoretic constructs degree centrality and path analysis are particularly exploited.

We illustrated our methodology with depressive disorder, and there are three major components of our results: 1) inflammation and depression, 2) circadian phenomena and depression, 3) noradrenergic aspects of depression. Varying amounts of research have been devoted to each of these components, but little (if any) has considered all three together. Our results do not constitute a discovery in the sense of something previously not noticed by anyone. However, in several respects they contribute to various aspects of depression that are currently incompletely understood and have not been extensively studied. Insight into the interrelationships among all these components may materially contribute to unraveling the underlying pathophysiology of depression, thus underpinning more effective treatment (and prevention).

Acknowledgments

This research was supported in part by an appointment to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. This research was also supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

The first author also gratefully acknowledges funding from the Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (Cimbi.org), Otto Mønsteds Fond, Kaj og Hermilla Ostenfelds Fond, and the Ingeniør Alexandre Haynman og hustru Nina Haynmans Fond.

References

1. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.
2. Swanson DR and Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, 1997.
3. Hristovski D, Stare J, Peterlin B, and Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. In *Studies in Health Technology and Informatics*, pages 1344–1348. IOS Press, 2001.
4. Weeber M, Kors JA, and Mons B. Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*, 6(3):277–286, 2005.
5. Hristovski D, Friedman C, Rindflesch TC, and Peterlin B. Exploiting semantic relations for literature-based discovery. In *AMIA Annual Symposium Proceedings*, volume 2006, pages 349–353, 2006.
6. Cohen T, Whitfield GK, Schvaneveldt RW, Mukund K, and Rindflesch T. EpiphaNet: an interactive tool to support biomedical discoveries. *Journal of Biomedical Discovery and Collaboration*, 5:22–44, 2010.
7. Ahlers CB, Hristovski D, Kilicoglu H, and Rindflesch TC. Using the literature-based discovery paradigm to investigate drug mechanisms. In *AMIA Annual Symposium Proceedings*, pages 6–10, 2007.
8. Hristovski D, Kastrin A, Peterlin B, and Rindflesch T. Combining semantic relations and DNA microarray data for novel hypotheses generation. *Linking Literature, Information, and Knowledge for Biology*, pages 53–61, 2010.
9. Smalheiser NR, Torvik VI, Bischoff-Grethe A, Burhans LB, Gabriel M, Homayouni R, Kashef A, Martone ME,

Perkins GA, Price DL, et al. Collaborative development of the arrowsmith two node search interface designed for laboratory investigators. *Journal of Biomedical Discovery and Collaboration*, 1(1):8, 2006.
10. Dudley JT, Schadt E, Sirota M, Butte AJ, and Ashley E. Drug discovery in a multidimensional world: Systems, patterns, and networks. *Journal of Cardiovascular Translational Research*, 3:438–47, 2010.
11. Rindflesch TC and Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003.
12. Aronson AR and Lang FM. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–36, 2010.
13. Ahlers CB, Fiszman M, Demner-Fushman D, Lang F, and Rindflesch TC. Extracting semantic predications from MEDLINE citations for pharmacogenomics. In *Pacific Symposium on Biocomputing 2007*, pages 209–220, 2007.
14. Freeman LC. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.
15. Vogelstein B, Lane D, Levine AJ, et al. Surfing the p53 network. *Nature*, 408(6810):307–310, 2000.
16. Anyanwu K, Maduko A, and Sheth A. SemRank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th International Conference on World Wide Web*, pages 117–127, 2005.
17. Dupont P, Callut J, Dooms G, Monette JN, and Deville Y. Relevant subgraph extraction from random walks in a graph. *Research Report RR*, 380167, 2006.
18. Hagberg AA, Schult DA, and Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, 2008.
19. Tarjan R. Depth-first search and linear graph algorithms. In *Conference Record 1971 Twelfth Annual Symposium on Switching and Automata Theory*, pages 114–121, 1971.
20. Kennaway DJ. Review: Clock genes at the heart of depression. *Journal of Psychopharmacology*, 24(2 suppl):5–14, 2010.
21. Stehle JH, von Gall C, and Korf HW. Organisation of the circadian system in melatonin-proficient C3H and melatonin-deficient C57BL mice: a comparative investigation. *Cell and Tissue Research*, 309(1):173–182, 2002.
22. Lincoln G, Messager S, Andersson H, and Hazlerigg D. Temporal expression of seven clock genes in the suprachiasmatic nucleus and the pars tuberalis of the sheep: evidence for an internal coincidence timer. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13890–5, 2002.
23. Pang CS, Mulnier C, Pang SF, and Yang JCS. Effects of halothane, pentobarbital and ketamine on serum melatonin levels in the early scotophase in New Zealand white rabbits. *Neurosignals*, 10(5):310–316, 2001.
24. Lahiri S, Singh P, Singh S, Rasheed N, Palit G, and Pant KK. Melatonin protects against experimental reflux esophagitis. *Journal of Pineal Research*, 46(2):207–213, 2009.
25. Jiménez-Caliani AJ, Jiménez-Jorge S, Molinero P, Guerrero JM, Fernández-Santos JM, Martín-Lacave I, and Osuna C. Dual effect of melatonin as proinflammatory and antioxidant in collagen-induced arthritis in rats. *Journal of Pineal Research*, 38(2):93–99, 2005.
26. Chen WQ, Zhao XL, Hou Y, Li ST, Hong Y, Wang DL, and Cheng YY. Protective effects of green tea polyphenols on cognitive impairments induced by psychological stress in rats. *Behavioural Brain Research*, 202(1):71–76, 2009.
27. Sirivelu MP, Shin AC, Perez GI, MohanKumar PS, and MohanKumar SMJ. Effect of l-dopa on interleukin-1-induced suppression of luteinizing hormone secretion in intact female rats. *Human Reproduction*, 24(3):718–725, 2009.
28. Sirivelu MP, Burnett R, Shin AC, Kim C, MohanKumar PS, and MohanKumar SMJ. Interaction between GABA and norepinephrine in interleukin-1beta-induced suppression of the luteinizing hormone surge. *Brain Research*, 1248:107–14, 2009.
29. MohanKumar SMJ, Smith CL, and MohanKumar PS. Central adaptation to chronic administration of interleukin-1beta (IL-1beta) in rats. *Brain Research Bulletin*, 62(1):71–6, 2003.
30. MohanKumar SMJ and MohanKumar PS. Systemic Interleukin-1 [beta] stimulates the simultaneous release of norepinephrine in the paraventricular nucleus and the median eminence. *Brain Research Bulletin*, 65(5):451–456, 2005.
31. Rosmaninho-Salgado J, Araújo IM, Álvaro AR, Mendes AF, Ferreira L, Grouzmann E, Mota A, Duarte E, et al. Regulation of catecholamine release and tyrosine hydroxylase in human adrenal chromaffin cells by interleukin-

1β: role of neuropeptide Y and nitric oxide. *Journal of Neurochemistry*, 109(3):911–922, 2009.
32. Rosenwasser AM. Circadian clock genes: Non-circadian roles in sleep, addiction, and psychiatric disorders? *Neuroscience & Biobehavioral Reviews*, 34(8):1249–1255, 2010.
33. Abad VC, Sarinas PSA, and Guilleminault C. Sleep and rheumatologic disorders. *Sleep Medicine Reviews*, 12(3):211–228, 2008.
34. Miller AH, Ancoli-Israel S, Bower JE, Capuron L, and Irwin MR. Neuroendocrine-immune mechanisms of behavioral comorbidities in patients with cancer. *Journal of Clinical Oncology*, 26(6):971, 2008.
35. Srinivasan V, Spence DW, Trakht I., Pandi-Perumal SR, Cardinali DP, and Maestroni GJ. Immunomodulation by melatonin: its significance for seasonally occurring diseases. *Neuroimmunomodulation*, 15(2):93–101, 2008.
36. Raison CL, Capuron L, and Miller AH. Cytokines sing the blues: inflammation and the pathogenesis of depression. *Trends in Immunology*, 27(1):24–31, 2006.
37. Anisman H. Cascading effects of stressors and inflammatory immune system activation: implications for major depressive disorder. *Journal of Psychiatry & Neuroscience: JPN*, 34(1):4, 2009.
38. Itoi K and Sugimoto N. The brainstem noradrenergic systems in stress, anxiety and depression. *Journal of Neuroendocrinology*, 22(5):355–361, 2010.
39. Kennedy SH and Rizvi SJ. Emerging drugs for major depressive disorder. *Expert Opinion on Emerging Drugs*, 14(3):439–53, 2009.
40. El Mansari M, Guiard BP, Chernoloz O, Ghanbari R, Katz N, and Blier P. Relevance of norepinephrine–dopamine interactions in the treatment of major depressive disorder. *CNS Neuroscience & Therapeutics*, 16(3):e1–e17, 2010.
41. De Berardis D, Conti CM, Serroni N, Moschetta FS, Olivieri L, Carano A, Salerno RM, Cavuto M, Farina B, Alessandrini M, et al. The effect of newer serotonin-noradrenalin antidepressants on cytokine production: a review of the current literature. *International Journal of Immunopathology and Pharmacology*, 23(2):417–22, 2010.
42. Nichols AI, Tourian KA, Tse SY, and Paul J. Desvenlafaxine for major depressive disorder: incremental clinical benefits from a second-generation serotonin-norepinephrine reuptake inhibitor. *Expert Opinion on Drug Metabolism & Toxicology*, 6(12):1565–74, 2010.
43. Dunn AJ, Wang J, and Ando T. Effects of cytokines on cerebral neurotransmission. *Cytokines, Stress, and Depression*, pages 117–127, 1999.
44. Irwin M, Clark C, Kennedy B, Christian GJ, and Ziegler M. Nocturnal catecholamines and immune function in insomniacs, depressed patients, and control subjects. *Brain, Behavior, and Immunity*, 17(5):365–372, 2003.
45. Rief W, Mills PJ, Ancoli-Israel S, Ziegler MG, Pung MA, and Dimsdale JE. Overnight changes of immune parameters and catecholamines are associated with mood and stress. *Psychosomatic Medicine*, 72(8):755, 2010.

Appendix D

MEDLINE MeSH indexing: lessons learned from machine learning and future directions

---

Antonio Jimeno-Yepes, Bartłomiej Wilkowski, James G. Mork, Elizabeth Van Lenten, Dina Demner Fushman, Alan R. Aronson. MEDLINE MeSH indexing: lessons learned from machine learning and future directions. *ACM SIGHIT International Health Informatics Symposium*, 2011, (pp. 5). Accepted.

MEDLINE MeSH indexing: lessons learned from machine learning and future directions

Antonio Jimeno-Yepes*, James G. Mork*, Bartłomiej Wilkowski†, Dina Demner-Fushman* and Alan R. Aronson*
*National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
Email: antonio.jimeno@gmail.com; {mork, ddemner, alan}@nlm.nih.gov
†Technical University of Denmark
DTU Informatics
Richard Petersens Plads
B321, DK-2800, Kongens Lyngby, Denmark
Email: wilkowskib@gmail.com

***Abstract*—Due to the large yearly growth of MEDLINE, MeSH indexing is becoming a more difficult task for a relatively small group of highly qualified indexing staff at the US National Library of Medicine (NLM). The Medical Text Indexer (MTI) is a support tool for assisting indexers; this tool relies on MetaMap and a k-NN approach called PubMed Related Citations (PRC). Our motivation is to improve the quality of MTI based on statistical learning. Typical statistical learning approaches fit this indexing task into text categorization. In this work, we have studied some Medical Subject Headings (MeSH) recommended by MTI and analyzed the issues when using standard machine learning algorithms. We show that in some cases statistical learning can improve the annotations already recommended by MTI, that statistical learning based on low variance methods achieves better performance and that each MeSH heading presents a different behavior. In addition, there are several factors which make this task difficult (e.g. limited access to the full-text of the citations) which provide direction for future work.**

***Keywords*-MeSH indexing, MEDLINE, text categorization, machine learning**

I. Introduction

MEDLINE® citations are indexed using the Medical Subject Headings (MeSH)® controlled vocabulary. This indexing is performed by a relatively small group of highly qualified indexing staff at the US National Library of Medicine (NLM). Their task is becoming more difficult due to the ever increasing size of MEDLINE, currently around 700k articles per year[1]. We hope that the situation can be eased through improvements to the recommendations made by NLM's indexing tool, the Medical Text Indexer (MTI) [1], [2].

MTI is a support tool for assisting indexers as they add MeSH indexing to MEDLINE. MTI has two main components: MetaMap [3] and the PubMed® Related Citations (PRC) algorithm. MetaMap performs an analysis of the citations and annotates them with Unified Medical Language System (UMLS)® concepts. Then, the mapping from UMLS to MeSH follows the *Restrict-to-MeSH* [4] approach which is based primarily on the semantic relationships among UMLS concepts. The PRC [5] algorithm is a modified k-NN algorithm which relies on document similarity to assign MeSH headings (MHs). This method attempts to increase the recall of MetaMap by proposing indexing candidates for MeSH headings which are not explicitly present in the citation but which are used in similar context.

Our motivation is to improve MTI's recommendations using statistical learning because there is a large number of MeSH headings, around 26k, and previously indexed citations are available as training data. On the other hand, indexers have access to the full-text. Automatic indexing has no access to this due to license restrictions.

We encounter issues, some of which are common to text categorization:

1) Imbalance between the number of positive and negative instances where the negative class usually overwhelms the positive one. Some machine learning algorithms have difficulty with this imbalance. We tested several approaches to deal with this issue to balance the datasets and to use a method based on the optimization of a multivariate measure instead of relying on accuracy. Joachims [6] proposed an adaptation of SVM to optimize measures like F-measure or the area under the ROC-curve instead of accuracy, being an alternative to balancing the positive and negative instances.
2) Even if a MeSH heading is correctly identified with a citation, it might not be significant enough to be included in the indexing.
3) Inconsistencies in the annotations might appear due to:

[1]http://www.nlm.nih.gov/bsd/bsd_key.html

a) Inconsistency in MeSH indexing [7].
b) Changes in indexing policy over time can introduce inconsistencies with previously-indexed citations. This can even apply to routine changes to the structure of MeSH. In the selection of our set we carefully avoided this issue by selecting MHs which were already in MeSH during the current indexing period.

In this paper, we study the use of machine learning algorithms in the task of MeSH indexing for some MeSH headings and present several characteristics of the task. We show that the citation text has limited prediction capability and that other sources of information (e.g. fulltext) or representations of the citations could still be explored. In the discussion, we point to future work and, based on statistics about MEDLINE indexing and MTI's performance, we suggest MHs to be considered as next study candidates.

II. Related work

Previous work has seen the indexing task as a text categorization task. The large body of related work provides valuable insights with respect to classification of MEDLINE citations and feature selection methods.

We find that most of the methods fit either into pattern matching methods which are based on a reference terminology (like UMLS or MeSH) and machine learning approaches which learn a model from examples of previously indexed citations.

Among the pattern matching methods we find the first component of MTI, as mentioned above, and an information retrieval approach by Ruch [8]; in Ruch's system the categories are the documents and the query is the text to be indexed. Pattern matching considers only the inner structure of the terms but not the terms with which they co-occur. This means that if an article is related to a MeSH heading but does not appear in the reference source (usually restricted to abstract text and title due to availability of full-text), it will not be suggested. Machine learning based on previously indexed citations might help to overcome this problem.

This problem has been approached in several ways from a machine learning point of view. Machine learning methods tend to be ineffective with a large number of categories. Small scale studies with machine learning approaches already exist [9], [10]. But the presence of a large number of categories has forced machine learning approaches to be combined with information retrieval methods designed to reduce the search space. For instance, PRC and a k-NN approach by Trieschnigg et al. [11] look for similar citations in MEDLINE and predict MeSH headings by a voting mechanism on the top-scoring citations. Experience with MTI shows that k-NN methods produce high recall but low precision indexing. Other machine learning algorithms have been evaluated which rely on a more complex representation of the citations which do not rely only on unigrams or bigrams, e.g., learning based on ILP (Inductive Logic Programming) [12].

III. Machine learning analysis

Experiments have been performed on the MTI experiment set for the 2009 MEDLINE indexing. This set-up allows avoiding any interference provided by policy change in the indexing. We have selected candidate MHs highly represented in MEDLINE but with poor recall performance by MTI. The list of selected MHs is found in Table I along with their MeSH identifiers and tree code[2]. MTI performance for each MH is available in Table IV.

MeSH Heading	Unique ID	Tree Number
Acute Disease	D000208	C23.550.291.125
Gene Expression	D015870	G05.355.310
Health Services	D006296	N02.421
Hormones	D006728	D06.472/D27.505.696.399.472
Infection	D007239	C01.539
RTPCR	D020133	E05.393.620.500.725

Table I
Selected MeSH headings based on 2010 MeSH

This selection has been previously used in [13]. In the current work a two stage approach to the problem is presented, in which the first step attempts to improve recall while the latter to increase precision. We focus on a deeper analysis of the second step, in which a previously selected subset of documents is further analyzed according to the methods and the representation of the documents.

In the first step, the idea is to reduce the whole dataset to ease the work with statistical learning algorithms. This reduction is performed by reducing the feature space using Latent Dirichlet Allocation (LDA) [14] to extract the most salient terms in the groups and selecting the terms with a higher prediction performance based on the combination of decision trees (DT) common branches on cross-validation sets and decision trees. The DT derived rules (recall rules) reduce the total set of citations to be considered by the false positive filtering study, see Table IV. We can see that in almost all the cases we can reduce the size of the set, keeping recall high for each MeSH heading but still with low precision.

In Table II, we show several terms which appeared in the LDA analysis for *Gene Expression*. We find that terms like *expression* have high coverage but low precision, since there are terms which can be used in different situations. On the other hand, we find the term *gene expression* which has lower recall, but surprisingly the precision is still very low. This means that there are cases in which the term *gene expression* appears in the citation but does not qualify to be included as a candidate MH. Machine learning will not only

[2]RTPCR stands for Reverse Transcriptase Polymerase Chain Reaction

have to ensure that the term is used in the proper sense but that it is significant enough to qualify, showing further the complexity of the task.

Term	Rec	Prec	F1
gene expression	0.2543	0.1668	0.2014
mrna	0.2965	0.1243	0.1752
expression	0.7704	0.0933	0.1664
gene	0.5492	0.0725	0.1281
expressed	0.3033	0.0771	0.1230

Table II
GENE EXPRESSION FEATURE PREDICTION STUDY

Some of the MeSH headings in our study are parents of more specific headings in the MeSH taxonomy (e.g. *Hormones*). These more specific headings (e.g. *thyroid hormones*) might be used for indexing instead of the MHs we are considering. To evaluate the impact of this phenomenon we have identified the children of the MHs under study. In Table III we show that some MHs like *Hormones* and *Infection* have a large number of children and seem to overlap with the indexing performed for these MHs (FP+Children). In the case of *Hormones*, half of the false positives (FP) are indexed with a hormone type. Methods based on pattern matching might avoid this issue selecting the MH matching the largest span of text. Examples of these methods are MetaMap and Ruch's approach.

MeSH Heading	Children	FP+Children	Total FP
Gene Expression	3	984	24978
Health Services	2	76	27475
Hormones	212	2290	4181
Infection	148	3408	49796

Table III
OVERLAP OF FPS AND ANNOTATION OF MORE SPECIFIC MESH HEADINGS

To the reduced set produced by the recall rules, we have applied the following machine learning algorithms. Each algorithm relies on different learning bias which would allow closer examination of the results for each one of the cases.

1) Traditional classifiers (SVM, Naïve Bayes, decision trees and k-NN).
2) Ensemble of classifiers, in some cases can reduce the variance of decision trees or can consider complementary views of the problem by different learning algorithms (boosting, bagging, voting; e.g. AdaBoost).
3) SVM with multivariate measures [6]

False positive filtering experiments (Filtering) have been performed for each one of the learning algorithms listed above. Unigrams and bigrams are used in the representation of the documents. Results are presented in Table IV. We show the MTI results, MTI with machine learning filtering (MTI+Filtering), the outcome of the recall analysis and the recall analysis with machine learning filtering (Recall+Filtering).

The machine learning sets used are the MTI set and the reduced set from the recall analysis. The result, as observed already in [13], is that machine learning improves the precision of the MeSH heading recommendation but at the cost of recall.

We also show results of the children analysis in Table IV for *Hormones* and *Infection*. We can see that children analysis improves the performance of the recommendations, meaning that the MeSH structure should be further studied in order to improve the recommendations.

From the machine learning algorithms used in the experiments, AdaBoost, SVMs and multivariate SVM achieve the best performance in many of the filtering results, meaning that low variance methods achieve a more interesting performance. On the other hand, decision trees achieve the lowest performance which correlates with previous studies on text categorization.

Acute Disease	Prec	Rec	F1	F2
MTI	0.2664	0.1580	0.1984	0.1720
MTI+Filtering	**0.4272**	0.1395	0.2103	0.1612
Recall analysis	0.1176	**0.8562**	0.2068	0.3795
Recall+Filtering	0.1941	0.6611	**0.3001**	**0.4463**
Gene Expression	Prec	Rec	F1	F2
MTI	0.1958	0.2712	**0.2274**	0.2518
MTI+Filtering	**0.2642**	0.1389	0.1896	0.1805
Recall analysis	0.0645	**0.8165**	0.1195	0.2450
Recall+Filtering	0.1130	0.5220	0.1858	**0.3029**
Health Services	Prec	Rec	F1	F2
MTI	0.1810	0.3533	**0.2394**	**0.2968**
MTI+Filtering	**0.2376**	0.2173	0.2270	0.2211
Recall analysis	0.0169	**0.6293**	0.0329	0.0763
Recall+Filtering	0.0723	0.3547	0.1201	0.1992
Hormones	Prec	Rec	F1	F2
MTI	0.0726	0.4000	0.1229	0.2103
MTI+Filtering	**0.1310**	0.2800	**0.1785**	**0.2281**
Recall analysis	0.0328	**0.6311**	0.0624	0.1359
Recall+Filtering	0.0839	0.3600	0.1361	0.2172
Recall no children	0.0698	0.6311	0.1258	0.2421
Recall nc filter	0.1845	0.3911	0.2507	0.3195
Infection	Prec	Rec	F1	F2
MTI	0.0649	0.4013	0.1117	0.1970
MTI+Filtering	**0.1568**	0.2492	**0.1925**	**0.2229**
Recall analysis	0.0048	**0.7767**	0.0095	0.0234
Recall+Filtering	0.0216	0.4660	0.0412	0.0910
Recall no children	0.0051	0.7767	0.0102	0.0251
Recall nc filter	0.0276	0.4854	0.0523	0.1126
RTPCR	Prec	Rec	F1	F2
MTI	0.2790	0.3738	0.3213	0.3535
MTI+Filtering	**0.4267**	0.2844	**0.3413**	0.3047
Recall analysis	0.0931	**0.7191**	0.1648	0.3066
Recall+Filtering	0.2048	0.4863	0.2883	**0.3815**

Table IV
PRECISION ANALYSIS RESULT

IV. Discussion

In our study, we have used a data set from 2009 MTI experiments, and we have analyzed some of the characteristics of the results obtained by applying machine learning on them. We have presented the issues which machine learning algorithms face when dealing with MeSH indexing.

As we have noted above, each MH seems to have a different behavior according to the method used. Since there are 26k MHs, to train and maintain up-to-date a system which can manage the different MHs, it might be possible to place the effort on highly represented MHs. Systems based on k-NN [5], [11] or matching strategies like MetaMap and Ruch's approach [8] manage the size problem efficiently. In this section, we present different statistics on the MeSH indexing which could help deciding on focusing the effort on a specific set of MHs.

Table V shows the micro/macro-average performance of MTI. We can see that while recall is almost the same, precision is much lower for micro-average. This might mean that there are MHs which are highly represented in MEDLINE indexing (e.g. Female) for which MTI achieves a result with low precision.

	Precision	Recall	F-measure
Macro-average	0.4164	0.5111	0.4589
Micro-average	0.3268	0.5118	0.3989

Table V
MTI micro averaging based on $\ln frequency$

Table VI shows the distribution of MHs according to their occurrence frequency in MEDLINE. In order to properly distribute the MHs, we have placed them into bins according to the logarithm of the frequency. MHs indicate the number of individual MHs, the total is the actual total mention of MHs, and precision, recall and F-measure is the average performance in each one of these categories. MTI's performance seems to decrease slightly as the total number of citations indexed by the MHs increases. The exception is the last category with only the single MH *Humans*. We can see that the last five categories have a low number of MHs but the total number of occurrences in MEDLINE is quite high. The most popular terms in our dataset are *Humans* with 471,467 occurrences, *Female* with 233,499 and *Male* with 227,052.

There are MHs with very low number of mentions in MEDLINE. We can assume that these MHs are rare, but even if you find the term it does not mean that it is significant enough to be added to the indexing.

We find as well that there are 1,314 MHs which are never considered for indexing[3]. Some MHs are used to specify the *Publication Characteristics* (Tree V), which in some cases allow the identification of funding support for the article[4]. Other MHs are used to organize the MeSH taxonomy.

$\ln(freq)$	MHs	Total	Prec	Rec	F1
0	833	833	0.2878	0.4898	0.3626
1	1933	5704	0.4448	0.5108	0.4755
2	3375	27296	0.4910	0.5363	0.5126
3	4393	94692	0.4834	0.5430	0.5115
4	4795	273297	0.4671	0.5456	0.5033
5	4313	650906	0.4230	0.5399	0.4743
6	2698	1091380	0.3860	0.5454	0.4520
7	1319	1392237	0.3500	0.5602	0.4309
8	465	1303683	0.3321	0.5574	0.4162
9	115	898067	0.3263	0.5208	0.4012
10	22	429109	0.4074	0.4413	0.4237
11	7	369217	0.4735	0.3472	0.4007
12	5	874276	0.5817	0.2964	0.3927
13	1	471467	0.9155	0.6914	0.7878

Table VI
MTI macro averaging based on $\ln frequency$

Table VII shows the macro average performance of MTI according to each one of the MeSH trees. A detailed list of the current tree codes is available from[5]. We can see that there are trees which contain a low number of MeSH headings but embody a large number of indexed citations like CT (Check Tags), G (Analytical, Diagnostic and Therapeutic Techniques and Equipment) and E (Phenomena and Processes).

One possible next step would consist of focusing on these sets of MeSH headings and try, in addition, to identify commonalities among the MHs.

Tree	MHs	Total	Prec	Rec	F1
A	1614	480326	0.3723	0.5404	0.4641
B	3546	248804	0.5459	0.6465	0.5989
C	4394	757400	0.4600	0.5682	0.5107
CT	34	1804516	0.4393	0.3007	0.3041
D	8805	1287185	0.4323	0.5327	0.4740
E	2396	1412951	0.3515	0.4146	0.3590
F	739	281784	0.3208	0.3944	0.3352
G	1360	822398	0.2970	0.4253	0.3491
H	292	91239	0.2796	0.3122	0.2656
I	410	133224	0.3068	0.3389	0.2977
J	193	46574	0.3123	0.4165	0.3557
K	145	13341	0.3135	0.2505	0.2404
L	246	86219	0.2171	0.2519	0.2066
M	154	48106	0.3371	0.3564	0.3106
N	737	233104	0.2528	0.2536	0.2194
V	146	0	0.0000	0.0000	0.0000
Z	377	134993	0.5006	0.5391	0.5125

Table VII
MTI macro averaging based on MeSH Tree code

[3]From the MEDLINE Baseline http://mbr.nlm.nih.gov/index.shtml

[4]http://www.nlm.nih.gov/bsd/funding_support.html

[5]http://www.nlm.nih.gov/mesh/trees.html

We have performed experiments on the text provided by the abstract and title of the citations. The results point out that the citations might not provide enough information to index the citations, e.g. for around 15% of the citations only the title is present. Further studies on full-text might be required, but only 15% of the PMIDs in our dataset could be matched to full-text identifiers in PubMed Central®. Specific feature selection and combination might be required to process the articles efficiently.

Another possibility is to consider existing meta-data already available in the citations. One possibility is to correlate the MeSH headings with the journals in which the citations appears. Another possibility is to use the Journal Descriptor indexing which has already been proposed in the literature [15].

V. Conclusion

In this work, we have studied the use of different machine learning algorithms and seen that MeSH indexing suffers from similar issues as other text categorization tasks. Nevertheless, low variance methods seem to achieve better performance. In addition, different methods exhibit behavior depending on the MH, and the position of the MH in the MeSH taxonomy might be partly responsible. The current approach is derived from the citation text, but the possibility of using full-text is restricted. We propose to further study alternative representations of the citations.

Acknowledgment

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. This research was supported in part by an appointment to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. The third author also gratefully acknowledges funding from the Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (Cimbi.org), Otto Mønsteds Fond, Kaj og Hermilla Ostenfelds Fond, and the Ingeniør Alexandre Haynman og hustru Nina Haynmans Fond.

References

[1] A. Aronson, O. Bodenreider, H. Chang, S. Humphrey, J. Mork, S. Nelson, T. Rindflesch, and W. Wilbur, "The NLM Indexing Initiative." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000, p. 17.

[2] A. Aronson, J. Mork, C. Gay, S. Humphrey, and W. Rogers, "The NLM Indexing Initiative's Medical Text Indexer," in *Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004]*. OCSL Press, 2004, p. 268.

[3] A. Aronson and F. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, p. 229, 2010.

[4] K. Fung and O. Bodenreider, "Utilizing the UMLS for semantic mapping between terminologies." American Medical Informatics Association, 2005.

[5] J. Lin and W. Wilbur, "PubMed related articles: a probabilistic topic-based model for content similarity," *BMC bioinformatics*, vol. 8, no. 1, p. 423, 2007.

[6] T. Joachims, "A support vector method for multivariate performance measures," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 377–384.

[7] M. Funk and C. Reid, "Indexing consistency in MEDLINE." *Bulletin of the Medical Library Association*, vol. 71, no. 2, p. 176, 1983.

[8] P. Ruch, "Automatic assignment of biomedical categories: toward a generic approach," *Bioinformatics*, vol. 22, no. 6, p. 658, 2006.

[9] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C. Aliferis, "Text categorization models for high-quality article retrieval in internal medicine," *Journal of the American Medical Informatics Association*, vol. 12, no. 2, pp. 207–216, 2005.

[10] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on MEDLINE document classification," in *AMIA Annual Symposium Proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 849.

[11] D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann, "MeSH Up: effective MeSH text classification for improved document retrieval," *Bioinformatics*, vol. 25, no. 11, p. 1412, 2009.

[12] A. Névéol, S. Shooshan, and V. Claveau, "Automatic inference of indexing rules for MEDLINE," *BMC bioinformatics*, vol. 9, no. Suppl 11, p. S11, 2008.

[13] A. Jimeno-Yepes, B. Wilkowski, J. Mork, E. Van Lenten, D. Demner Fushman, and A. Aronson, "A bottom-up approach to MEDLINE indexing recommendations," in *AMIA Symposium (submitted)*. American Medical Informatics Association, 2011.

[14] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] S. Humphrey, "Automatic indexing of documents from journal descriptors: A preliminary investigation," *Journal of the American Society for Information Science*, vol. 50, no. 8, pp. 661–674, 1999.

Appendix E

A bottom-up approach to MEDLINE indexing recommendations

Antonio Jimeno-Yepes, Bartłomiej Wilkowski, James G. Mork, Elizabeth Van Lenten, Dina Demner Fushman, Alan R. Aronson. A bottom-up approach to MEDLINE indexing recommendations. *American Medical Informatics Association Annual Symposium*, 1583–1592, Washington D.C., 2011. Published.

A bottom-up approach to MEDLINE indexing recommendations

Antonio Jimeno-Yepes, PhD[1], Bartłomiej Wilkowski, MS[2], James G. Mork, MS[1], Elizabeth Van Lenten, PhD[1], Dina Demner Fushman, MD, PhD[1], Alan R. Aronson, PhD[1]
[1] National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
[2] Technical University of Denmark, DTU Informatics, Richard Petersens Plads, B321, DK-2800, Kongens Lyngby, Denmark

Abstract

MEDLINE indexing performed by the US National Library of Medicine staff describes the essence of a biomedical publication in about 14 Medical Subject Headings (MeSH). Since 2002, this task is assisted by the Medical Text Indexer (MTI) program. We present a bottom-up approach to MEDLINE indexing in which the abstract is searched for indicators for a specific MeSH recommendation in a two-step process. In the first step, a rule-based triage significantly reduces the number of candidate citations to which the MeSH heading is recommended. In the second step, the candidate citation list is further reduced using supervised machine learning. Supervised machine learning combined with triage rules improves sensitivity of recommendations while keeping the number of recommended terms relatively small. Improvement in recommendations observed in this work warrants further exploration of this approach to MTI recommendations on a larger set of MeSH headings.

Introduction

Maintaining the quality of MEDLINE® indexing is made difficult by the demand of the ever increasing size of the biomedical literature on a relatively small group of highly qualified indexing staff at the US National Library of Medicine (NLM). We hope that the situation can be eased through improvements to the recommendations made by NLM's indexing tool, the Medical Text Indexer (MTI) [1,2].

MTI is a support tool for assisting indexers as they add MeSH® indexing to MEDLINE. MTI has two main components: MetaMap and the PubMed® Related Citations (PRC) algorithm. MetaMap performs an analysis of the citations and annotates them with Unified Medical Language System (UMLS)® concepts. Then, the mapping from UMLS to MeSH follows the *Restrict-to-MeSH* [3] approach which is based primarily on the semantic relationships among UMLS concepts. The PRC [4] algorithm is a modified k-NN algorithm which relies on document similarity to assign MeSH headings. This method intends to increase the recall of MetaMap by proposing indexing candidates for MeSH headings which are not explicitly present in the citation but have a similar context.

The NLM indexing process involves analysis of journal articles for subject matter and subsequent assignment of appropriate subject headings, drawn from MeSH, the NLM controlled vocabulary. There are 25,588 descriptors or main headings (MHs) in 2010 MeSH from which MTI recommends about 25 terms, on average. Based on the results for 142,262 citations processed by MTI between November 23, 2009 and February 8, 2010, the number of MHs we need to review for possible improvements in MTI recommendations is significantly smaller than 25,588. Figure 1 illustrates the breakout of the different MHs that can be removed from consideration. There are 12,350 MHs in the "B" (*Organisms*) and "D" (*Chemicals and Drugs*) MeSH trees (recommended automatically if the terms are found in the title), 1,854 MHs for which MTI recommendations using the current top-down approach are fairly accurate (precision over 60%), 251 MHs that are used for cataloging and other purposes but not for indexing, 3,609 MHs occurring less than 500 times in MEDLINE, and 832 MHs that are too general in nature. The remaining 6,692 (26.15% of the 2010 MeSH) need improved recommendations.

Our previous attempts to improve the quality of recommendations [5,6] indicate that the current, top-down method might be approaching the upper bound on its performance, and other methods need to be explored to improve recommendations for the remaining 26% of the headings.

The motivation for this work comes from an approach suggested by indexers who use certain *indicators* in the articles that lead to assignment of specific indexing terms which might complement MTI annotation. We describe a semi-automatic procedure we followed to identify triage rules and a filtering step which are designed to emulate the approach

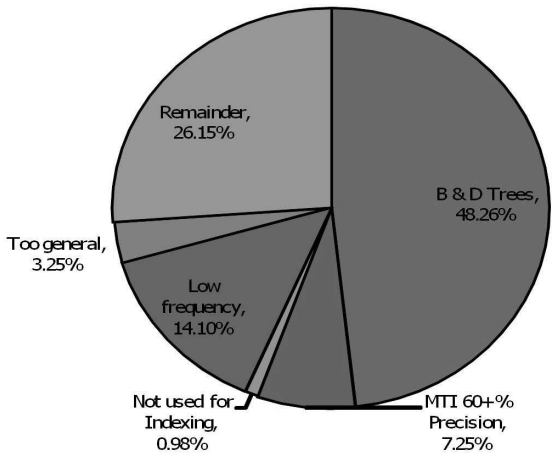


Figure 1: MTI accuracy primarily needs to be improved for 26% of MeSH suggestions

used by the indexers. We start our exploration of this bottom-up approach using rules developed by a domain expert for recommending the Carbohydrate Sequence heading. Then, we further evaluate the proposed approach on two MeSH headings from the latest MTI test set. The results encourage the exploration of this method with other MeSH headings.

Related work

Publication of the OHSUMED collection[7] containing all MEDLINE citations into 270 medical journals over a five-year period (1987-1991) including MeSH indexing, provided for a large body of data that enabled us to view MH assignment as a classification problem. The scope of the collection determines the subset of MeSH that can be explored. For example, Lewis et al.[8] and Ruiz and Srinivasan[9] used 49 categories related to heart diseases with at least 75 training documents, and Yetisgen-Yildiz and Pratt[10] expanded the number of headings to 634 disease categories. Poulter[11] provides an overview of these and other studies of classification methods applied to MEDLINE and MeSH subsets.

The two-step approach to document triage and filtering was implied in the definition of the Text REtrieval Conference (TREC) Genomics track 2004 and 2005 categorization tasks, in which the main task was to consider each document for routing for further expert review or not, and in the subtasks the documents were annotated with specific terms[12,13].

A growing body of work approaches retrieval of MEDLINE citations as a classification task. For example, MScanner classifies all MEDLINE citations as relevant to a set of positive examples submitted by a user or not[11], and Kastrin et al.[14] determine the likelihood of MEDLINE citations topical relevance to genetics research. The large body of related work provides valuable insights with respect to classification of MEDLINE citations and feature selection methods. Applicability of these methods and suitability of the features for our specific task of improving indexing suggestions needs to be explored further. We find that most of the methods fit either into pattern matching methods which are based on a reference terminology (like UMLS or MeSH) and machine learning approaches which learn a model from examples of previously indexed citations.

Among the pattern matching methods we find the first component of MTI, as shown above, and an information retrieval approach by Ruch[15]; in his system the categories are the documents and the query is the text to be indexed. Pattern

matching considers only the inner structure of the terms but not the terms with which they co-occur. This means that if a document is related to a MeSH heading but does not appear in the reference source, it will not be suggested. Machine learning based on previously indexed citations might help to overcome this problem.

This problem has been approached in several ways from a machine learning point of view. Machine learning methods tend to be ineffective with a large number of categories; MeSH contains more than 25k. Small scale studies with machine learning approaches already exist [16,10]. But the presence of a large number of categories has forced machine learning approaches to be combined with information retrieval methods designed to reduce the search space. For instance, PRC and a k-NN approach by Trieschnigg et al. [17] look for similar citations in MEDLINE and predict MeSH headings by a voting mechanism on the top-scoring citations. Experience with MTI shows that k-NN methods produce high recall but low precision indexing.

Other machine learning algorithms have been evaluated which rely on a more complex representation of the citations which do not only rely on unigrams or bigrams, e.g., learning based on ILP (Inductive Logic Programming) [18].

Methods

In our work, we intend to improve MTI's MEDLINE MeSH recommendations by targeting ones where MTI's performance is poor. In the first part of this section, we present the work performed by a domain expert building triage rules to recommend the Carbohydrate Sequence heading. Then, we present the bottom-up approach proposed in this paper which is composed of two methods. In the first method, triage rules related to the MeSH term under study are identified in a semi-automatic fashion. In the second method, a false positive filter is applied based on statistical learning algorithms.

Triage rules for recommending the Carbohydrate Sequence heading

The abstracts of the scientific publications are reviewed to identify strings potentially containing carbohydrate sequences. The following carbohydrate names are used to identify candidate strings: *GlcpNAc, GalpNAc, GalNAc, GlcNAc, Neu5Ac, NeuAc, GalpA, GlcpA, Galp, Glcp, Rhap, NANA, Man, Fuc, Gal, Glc*. Based on empirical results, the first five names are converted to lower case, and for the rest of the list case information is preserved. When one of the carbohydrate names (starting with the longest) is found, the extent of the continuous string of text (with no blanks) enclosing the name is identified. The string containing the name is searched for the remaining carbohydrates if it is longer than 4 characters. The occurrences of carbohydrates in the string are marked as found and counted. If at least three carbohydrates names (not necessarily unique) occur within the string and at least one of the names is longer than 3 characters (or the string contains digits or parentheses in addition to 3-character long names), the string is considered a Carbohydrate Sequence and MTI recommends the heading. The 3 character carbohydrates are too commonly found in text to be allowed on their own without the support of digits or parentheses. Figure 2 illustrates the rules applied to an excerpt from an abstract (PMID 1368642). First, the longer carbohydrate *GlcNAc* was found. The extent of the continuous string (marked by arrows) was identified next, and then carbohydrates *Man*, and *Fuc* were identified in the string. Due to this combination of three carbohydrates, MTI recommended the Carbohydrate Sequence heading.

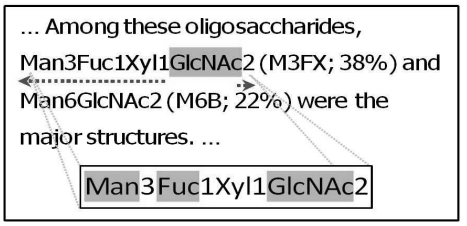


Figure 2: Excerpt detailing rules for identifying sequence

Semi-automatic learning of triage rules

Triage rules are learned in two steps: feature selection and rule learning, as shown in Figure 3. In the feature selection step, we the most salient terms from a set of training citations. In the rule learning step, we build models which target the MeSH Heading giving preference to high recall, at the cost of precision in many cases. The rules produced in this step will provide an upper bound on recall.

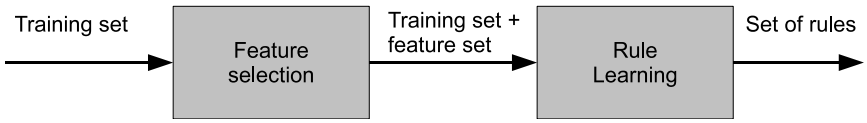


Figure 3: Triage rule learning

An example of triage rule derived from the Carbohydrate Sequence set is the following:

```
If any of the terms carbohydrate, polysaccharide, oligosaccharides combined
   with structure are found in citations in journals known to have Carbohydrate
Then assign Carbohydrate Sequence
```

Feature selection step

Feature selection is done before running the rule learning algorithm. Specifically, we used Mallet's implementation of Latent Dirichlet Allocation (LDA)[19] and Non-negative matrix factorization (NMF)[20]. From the derived model, we selected the top-n terms having a higher probability of belonging to a given topic. Examples of distilled terms for Carbohydrate Sequence are listed in Figure 5.

Rule induction step

Once the citations are represented by the most salient terms, we prepared a set of decision trees, which allow selecting the terms which correlate with the set of positive citations. The positive citations are the false negatives and the set of negative citations contain the false positive of the MTI predictions. Figure 4 shows a sample tree produced for *Carbohydrate Sequence* feature set. In this sample, the tree has been turned into a set of rules with output POS (recommend the MH) or NEG (do not recommend the MH). Once a rule is matched, the outcome is defined by the element after the colon punctuation mark. The rules are composed of the occurrence of a term in the citation and a combination of operators. The & sign is the Boolean operator AND and the ! sign is the negation operator. The second rule recommends Carbohydrate Sequence if the citation does not contain the terms *oligosaccharides polysaccharide* and *carbohydrate* but contains the term *gal*.

Filtering based on machine learning

The rule induction step improves recall but adds false positives in the process. In order to remove false positives and thereby improve precision, we applied machine learning. This task is considered a text categorization task; it attempts to filter out false positives by deciding if a given citation should be indexed with the give MeSH heading or not. We encounter issues, some of which are common to text categorization:

1. Imbalance between the number of positive and negative instances where the negative class usually overwhelms the positive one. Some machine learning algorithms have difficulty with this imbalance. We tested several approaches to deal with this issue: to balance the datasets, and to use a method based on the optimization of a

```
!oligosaccharides&!polysaccharide&!carbohydrate&!gal:NEG
!oligosaccharides&!polysaccharide&!carbohydrate&gal:POS
!oligosaccharides&polysaccharide&!structure&!intensity:POS
!oligosaccharides&polysaccharide&!structure&intensity:NEG
!oligosaccharides&polysaccharide&structure&!text:POS
!oligosaccharides&polysaccharide&structure&text:NEG
oligosaccharides&!structures&!galactopyranoside&!xylosyl:POS
oligosaccharides&!structures&!galactopyranoside&xylosyl:NEG
oligosaccharides&!structures&galactopyranoside&!enzyme:NEG
oligosaccharides&!structures&galactopyranoside&enzyme:POS
oligosaccharides&structures&!relies&!released:POS
oligosaccharides&structures&!relies&released:POS
oligosaccharides&structures&relies:NEG
```

Figure 4: MALLET decision tree learned for Carbohydrate Sequence

multivariate measure instead of relying on accuracy. Joachims[21] proposed an adaptation of SVM to optimize measures like F-measure or the area under the ROC-curve instead of accuracy, being an alternative to balancing the positive and negative instances.

2. Even if a term is correctly identified with a citation, it might not be significant enough to be included in the indexing.
3. Inconsistencies in the annotations might appear due to:
   (a) Inconsistency in MeSH indexing[22]. In machine learning terms, this is class label noise. Several existing techniques could be considered to overcome this problem. One of them[23] consists of selecting only documents for which a low level of discrepancy exists among classifiers. Then, the model is learned only from instances with a high level of confidence.
   (b) Changes in indexing policy over time can introduce inconsistencies with previously-indexed citations. This can even apply to routine changes to the structure of MeSH. In the selection of our set we carefully avoided this issue by selecting terms which were already in MeSH during the indexing period. In addition, the span of time considered is small enough to avoid most of the indexing policy changes which might have occurred.

Filtering experimental setup

In this section, configurations for our experiments are specified. These configurations can be combined to build different models. Unigram and bigram representation from the title and abstract of the MEDLINE citations are used as features. The classifiers considered for the experiments are:

1. Traditional classifiers (SVM, Naïve Bayes, decision trees and k-NN).
2. Ensemble of classifiers, which can reduce the variance of decision trees or can consider complementary views of the problem by different learning algorithms (Boosting, bagging, voting, ...).
3. SVM with multivariate measures[21].

Evaluation

The methodology presented above has been evaluated on two datasets. The first one is a screening of MEDLINE with the MeSH heading Carbohydrate Sequence. The second one is a subset of MEDLINE used routinely for testing changes to MTI's algorithm.

Carbohydrate Sequence set

The Carbohydrate Sequence dataset is a subset of the 2010 MEDLINE baseline which contains 18,502,916 MEDLINE citations. The dataset consists of 2,307 citations found by Carbohydrate Sequence triage rules developed by a domain expert. The 2,307 citations that have the Carbohydrate Sequence heading (1,212 positive examples) and those that do not (1,095 negative examples) were further split into the training set containing 80% of the positive and negative examples and the test set containing the remaining 20% of the citations.

We consider citations with the Carbohydrate Sequence MH to be in the positive class because we are primarily interested in finding all citations that should be recommended for assigning the heading. The 80-20 split was performed using publication dates of the citations with the most recent citations in the test set. This split imitates the real-life situation in which a method is developed with an existing set of data and then applied to and tested on a future set.

A second set is developed for exploration of additional triage rules for the Carbohydrate Sequence heading. In this set we used 16,781 citations having the heading but not found by the current triage rules. We use a matching stratified sample of citations without the MH that were published in the journals containing at least one citation with the Carbohydrate Sequence MH as negative examples. These sets were also split into training and test subsets following the 80-20 rule described above.

MTI test set

This set is a subset of the 2009 MEDLINE baseline as used by the MTI team for verifying changes to MTI. We selected candidate terms highly represented in MEDLINE but with poor recall performance by MTI. The list of selected terms is found in Table 1.

MeSH Heading	Unique ID	Tree Number
Acute Disease	D000208	C23.550.291.125
Gene Expression	D015870	G05.355.310

Table 1: Selected MeSH headings

This set is split into training and test sets based on the publication date field (DP field in PubMed). The citations from the first 8 months of 2009 are used for training and the final 4 months for test. In the training set there are 409,279 citations with a total of 343,504 citations with abstract and in the test set, there are 255,493 citations with a total of 214,064 citations with abstracts.

Results

Carbohydrate Sequence set

In the preliminary exploration of the triage rules presented in the methods section, we noticed that many candidate citations do not get the MH assigned, while the majority of the citations having the Carbohydrate Sequence heading are not categorized as candidates. These observations led to expansion of the approach in two directions as introduced in the methods section: (1) using supervised machine learning to improve precision of recommendations for candidate citations found by the original triage rules, and (2) generation of new rules to expand the candidate set with a subsequent application of supervised machine learning. The goal of the triage step is to reduce the number of irrelevant citations to be processed in the second step. The goal of the machine learning step is to improve precision without losing recall.

The original Carbohydrate Sequence triage rules reduce the size of the set to be considered for Carbohydrate Sequence to 0.012% of the full document set with 6.7% recall and 52.5% precision. In the subsequent machine learning step the Maximum Entropy classifier trained on 1, 2, 3, and 4 token sequences and cutoff confidence level set above 0.2 reduced the number of wrong recommendations (precision of 53.6%) with almost all correct recommendations (90% of them).

The topics built using the positive examples (citations having the MH but no sequence strings) contained many key phrases pertaining to analysis methods (for example: *NMR spectroscopy*, *mass spectrometry*, *methylation analysis*), model organism and chemical names, which we found to be too general to pertain only to studies of carbohydrate sequences. However, topic analysis also provided pertinent terms, which we combined with the rules generated by the MALLET Decision Tree classifier shown in Figure 4. The new triage rules select citations for further consideration combining the common segments of the positive rules in Figure 4 and the topic terms found by LDA as follows:

1. Rule 1: If any of the terms *carbohydrate*, *polysaccharide*, *oligosaccharides* combined with structure are found in citations in journals known to have Carbohydrate
2. Rule 2: If two or more of the 26 terms listed for Carbohydrate Sequence (in Figure 5) are found in citations in journals known to have Carbohydrate
3. Rule 3: If rules 1 and 2 apply to the citation

carbohydrate(s), disaccharide(s), Fuc, Gal, GalNAc, galacturonic acid, GlcNAc, glucopyranosyl, glucuronic acid, glycan(s), glycosidic linkages, glycosylation, hyaluronic acid, iduronic acid, lipopolysaccharide(s), LPS, Man, NeuAc, oligosaccharide(s), polysaccharide(s), sialyl Lewis, sialic acid, Smith degradation, sugar chains, triterpenoid saponins

Figure 5: Terms selected from topics built based on the positive training examples for Carbohydrate Sequence recommendation

The new rules were evaluated using the second set of citations that have the Carbohydrate Sequence MH and further reduce the set of candidate citations. Results are available in Table 2.

Rule	True Positives	False Positives	Precision	Recall	F-measure	F_2-measure
Rule 1	3108	6528	0.4761	0.1733	0.2541	0.1986
Rule 2	8144	40768	0.1998	0.4541	0.2775	0.3620
Rule 1 & 2	2391	3043	0.7857	0.1333	0.2280	0.1599
CH & structure	1234	3883	0.3178	0.0688	0.1131	0.0816
Poly & structure	1292	1712	0.7547	0.0720	0.1315	0.0880
Olig & structure	1233	1566	0.7874	0.0688	0.1265	0.0841

Table 2: Rules and precision/recall values for the training and test set

Compared to the results using the initial triage rules, we obtain different precision recall levels which in most of the cases are better than the original triage rules developed manually by a domain expert.

MTI test set

Triage rule learning

Table 3 shows the results of the recall analysis for the MTI set. As described in the methods section, decision trees were built from the selected features, and common sections were manually analyzed. After careful analysis of the feature sets and common trees, rules where manually selected for each MeSH heading in order to obtain high coverage of the MeSH headings with reasonable precision. In Table 3 the rules and the performance measures are shown. The recall values are significantly increased compared to baseline MTI performance, but this was negatively compensated by a noticeable decrease in precision values.

Filter analysis

In Table 4, we find the results from the precision analysis for each of the MeSH headings. In this table, we show only the result produced by the best performing learning algorithm. For each MeSH heading we show the MTI result, the

	Rule	True Positives	False Positives	Precision	Recall
Acute Disease	'acut'	1387	10409	0.1176	0.8562
Gene Expression	('protein' & 'express') | ('gene' & 'express') | ('cell' & 'express')	1722	24978	0.0645	0.8165

Table 3: Rules and precision/recall values for the training and test set

MTI result with filtering, the recall analysis result, and the recall result with filtering. Overall, we find that filtering improves the precision but at a high recall cost. F-measure and F_2-measure results vary according to the method.

Acute Disease	True Positives	False Positives	Precision	Recall	F-measure	F_2-measure
MTI	256	705	**0.2664**	0.1580	0.1984	0.1720
Filtering	226	303	**0.4272**	0.1395	0.2103	0.1612
Triage rules	1387	10409	0.1176	**0.8562**	0.2068	0.3795
Triage + filtering	1071	4447	0.1941	0.6611	**0.3001**	**0.4463**
Gene Expression	True Positives	False Positives	Precision	Recall	F-measure	F_2-measure
MTI	572	2349	**0.1958**	0.2712	**0.2274**	0.2518
Filtering	293	816	**0.2642**	0.1389	0.1896	0.1805
Triage rules	1722	24978	0.0645	**0.8165**	0.1195	0.2450
Recall + filtering	1101	8639	0.1130	0.5220	0.1858	**0.3029**

Table 4: Results comparing the different analyses

Discussion

Focusing on the Carbohydrate Sequence set, the original triage rules based on identification of carbohydrate names are too specific to capture all candidates for recommending the Carbohydrate Sequence MH. To our surprise, the rules also capture a significant number of citations that should not be recommended for this MH. Our domain expert reviewed five citations that our best scoring classifier erroneously assigned to the positive class with highest confidence and concluded that those citations qualify for the MH. This indicates that the actual accuracy of the original rule might be higher than in our evaluation. In both cases, the filtering and the semi-automatically derived triage rules achieve results appropriate for the MEDLINE MeSH indexing.

For the MTI set, we show that semi-automatic generation of triage rules is potentially helpful to MEDLINE MeSH indexing. From the MTI analysis, we find that only in the case of *Acute Disease* did the recall and filtering analysis provide a result much higher compared to MTI. We can see that precision can be largely improved using machine learning based filters. On the other hand, recall decreases significantly in most of the cases. The F-measure is sometimes improved due to the increase in precision but the F_2-measure is lower in almost all scenarios due to the loss in recall. We plan to evaluate this approach with a larger set of MeSH headings occurring in the MTI set used in the experiments, and not only on MeSH headings with MTI's poor performance.

In addition, considering triage rules in the MTI set, precision is much lower compared to the original MTI prediction even if recall is much higher. After filtering, we find that precision is largely improved compared to the results from the recall analysis. The precision values are lower compared to the results obtained with MTI, so in many cases the final F-measure is lower. But for the F_2-measure *Acute Disease* and *Gene Expression* achieve better performance. Only in the case of *Acute Disease* we find that the F-measure and F_2-measure are improved in all the scenarios.

If we consider the machine learning algorithms in the filtering step, we find that low variance methods have better performance in comparison to low bias and high variance methods. Examples of low variance methods are SVM and AdaBoost. In many learning algorithms, balancing the dataset produced an improvement in the performance of the machine learning algorithms. Multivariate optimization achieved the highest performance with *Acute Disease* while AdaBoost achieved the highest improvement with *Gene Expression*.

Conclusions and Future work

Our work confirms that the method presented in this paper produces improved recommendations of some MeSH headings compared to existing methods and to manual assessment in the case of Carbohydrate Sequence. We plan to explore the scalability of the proposed approach, applying it to other headings for which indexing recommendations need to be improved.

The results presented in this work have been produced using unigrams and bigrams from the title and abstract. Other features extracted from text, like the position of the MeSH heading in the document or normalization of the features based on MetaMap, could be combined to improve the current results. This might then require moving to full-text analysis.

It should be mentioned that our results are probably negatively affected by papers which a title but no abstract (around 17% of all papers in the dataset). The fact that NLM indexers use the full-text of an article to perform MEDLINE indexing further argues for an extension of the current study to the use of full-text, not just titles and abstracts. Further studies on full-text might be required, but only 15% of the PMIDs in our dataset could be matched to PMC identifiers. Specific feature selection and combination might be required to process the articles efficiently.

Acknowledgments

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. This research was supported in part by an appointment to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. The second author also gratefully acknowledge funding from the Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (Cimbi.org), Otto Mønsteds Fond, and the Ingeniør Alexandre Haynman og hustru Nina Haynmans Fond.

References

1. A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, JG Mork, SJ Nelson, TC Rindflesch, and WJ Wilbur. The NLM Indexing Initiative. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2000.
2. A.R. Aronson, J.G. Mork, C.W. Gay, S.M. Humphrey, and W.J. Rogers. The NLM Indexing Initiative's Medical Text Indexer. In *Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004]*, page 268. OCSL Press, 2004.
3. K.W. Fung and O. Bodenreider. Utilizing the UMLS for semantic mapping between terminologies. American Medical Informatics Association, 2005.
4. J. Lin and W.J. Wilbur. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):423, 2007.
5. C.W. Gay, M. Kayaalp, and A.R. Aronson. Semi-automatic indexing of full text biomedical articles. In *AMIA Annual Symposium Proceedings*, volume 2005, page 271. American Medical Informatics Association, 2005.
6. A.R. Aronson, J.G. Mork, A. Névéol, S.E. Shooshan, and D. Demner-Fushman. Methodology for creating UMLS content views appropriate for biomedical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2008, page 21. American Medical Informatics Association, 2008.
7. W. Hersh, C. Buckley, TJ Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201. Springer-Verlag New York, Inc., 1994.
8. D.D. Lewis, R.E. Schapire, J.P. Callan, and R. Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306. ACM, 1996.
9. M.E. Ruiz and P. Srinivasan. Hierarchical neural networks for text categorization (poster abstract). In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 281–282. ACM, 1999.

10. M. Yetisgen-Yildiz and W. Pratt. The effect of feature representation on MEDLINE document classification. In *AMIA Annual Symposium Proceedings*, volume 2005, page 849. American Medical Informatics Association, 2005.
11. G.L. Poulter, D.L. Rubin, R.B. Altman, and C. Seoighe. MScanner: a classifier for retrieving Medline citations. *BMC bioinformatics*, 9(1):108, 2008.
12. W. Hersh, R.T. Bhuptiraju, L. Ross, P. Johnson, A.M. Cohen, and D.F. Kraemer. TREC 2004 genomics track overview. In *Proc. of the 13th Text REtrieval Conference*. Citeseer, 2004.
13. W. Hersh, A. Cohen, J. Yang, R.T. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 genomics track overview. In *Proceedings of the fourteenth text retrieval conference (TREC 2005)*. Citeseer, 2005.
14. A. Kastrin, B. Peterlin, and D. Hristovski. Chi-square-based scoring function for categorization of MEDLINE citations. *Methods of Information in Medicine*, 48:10–3414, 2009.
15. P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658, 2006.
16. Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C.F. Aliferis. Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association*, 12(2):207–216, 2005.
17. D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412, 2009.
18. A. Névéol, S. Shooshan, and V. Claveau. Automatic inference of indexing rules for MEDLINE. *BMC bioinformatics*, 9(Suppl 11):S11, 2008.
19. D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
20. D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.
21. T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.
22. M.E. Funk and C.A. Reid. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176, 1983.
23. X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *Proceedings of the 20th International Conference on Machine Learning*, 2010.

Appendix F

Neuroscientific literature search based on the location coordinates in brain - BredeQuery plugin for SPM environment

---

Bartłomiej Wilkowski. Neuroscientific literature search based on the location coordinates in brain - BredeQuery plugin for SPM environment. *Presented at: 2nd INCF Congress of Neuroinformatics, Front. Neur. Conference Abstract: Neuroinformatics 2009*, Pilsen, 2009. Published.

EVENT ABSTRACT

Back to Event

Neuroscientific literature search based on the location coordinates in brain - BredeQuery plugin for SPM environment

Bartlomiej Wilkowski[1*]

[1] Technical University of Denmark, DTU Informatics, Denmark

Current research in neuroscience, specifically neuroimaging, is based on the functional localization paradigm. New, state-of-the-art scanning techniques allow to detect haemodynamic response (dynamically regulated blood flow in brain) relating directly to neuronal activity in brain, hence to define functional localization of specific human behavior. There is a vast amount of new studies and publications about various human behaviours which are being mapped to specific brain regions where the significant change in activity occurs. The BredeQuery software [1] proposed herein, enables the user to search for related neuroscientific literature which reports various phenomena in similar brain regions. The search is done in a coordinate-based (not ordinary keyword-based) way and can be performed directly from the SPM (Statistical Parametric Mapping) - software platform run in Matlab and used inter alia in analysis of brain images.

The BredeQuery plugin is written in Matlab and Java and it runs directly from SPM. The coordinate-based search for publications is offered in the plugin using the Brede Database [2] *http://neuro.imm.dtu.dk/services/brededatabase/*. The graphical user interface of the BredeQuery plugin can be seen in the Figure 1.

The activation coordinates are grabbed from the SPM results window. Since the Brede Database is based on Talairach space coordinates, the BredeQuery plugin offers two MNI-to-Talairach (MTT) transformations. The grabbed coordinates can be then transformed from MNI space to Talairach using Mathew Brett's piece-wise affine transformation (brett) or Jack Lancaster et al.'s transformation (with three subtypes: SPM, FSL and pooled). The Lancaster's MTT-SPM transformation is set as default in the BredeQuery plugin.

The grabbed and optionally transformed coordinates can then be used for querying the Brede Database. The query results (related publications) can be shown in the internal Matlab web browser or can be automatically exported to one of the following bibliographic formats: BibTeX, Reference Manager, RefWorks, EndNote. It should be mentioned that the user is also able to perform the Brede Database query manually, without using the grabbed coordinates. The plugin was released on March 12, 2009. It can be found on the SPM community list of toolboxes: *http://www.fil.ion.ucl.ac.uk/spm/ext/#BredeQuery/* or can be directly downloaded from: *http://neuroinf.imm.dtu.dk/BredeQuery/*. Since the Brede Database is a relatively small database, the future development of the plugin is directed towards automatic retrieval of the articles also from huge, updated databases like PubMed. Integration with SKEEPMED (Semantic KEyword Extraction Pipeline for MEdical Documents) [3] is planned. This pipeline can be used for extracting relevant keywords from abstracts, mapping to ontological terms and constructing logical queries to other databases.

References

1. Wilkowski B. et al. - Coordinate-based meta-analytic search for the SPM neuroimaging pipeline - The BredeQuery plugin for SPM5. Presented at: HEALTHINF 2009

2. Nielsen F.A. et al. - The Brede database: a small database for functional neuroimaging. NeuroImage. Volume 19., Elsevier (June 2003)

3. Wilkowski B. et al. - Bridging the gap between coordinate- and keyword- based search of neuroscientific databases by UMLS-assisted semantic keyword extraction. Presented at: OHBM 2009. *http://neuroinf.imm.dtu.dk/cgi-bin/SKEEPMED/interactive_skeepmed.pl*

Conference : Neuroinformatics 2009, Pilsen, Czech Republic, 6 Sep - 8 Sep, 2009.

Presentation Type : Oral Presentation

Topic : Infrastructural and portal services

Citation : Wilkowski B (2009). Neuroscientific literature search based on the location coordinates in brain - BredeQuery plugin for SPM environment. *Front. Neur. Conference Abstract: Neuroinformatics 2009.* doi: 10.3389/conf.neuro.11.2009.08.071

Received : 22 May 2009; **Published Online:** 22 May 2009.

*** Correspondence :** Bartlomiej Wilkowski, Technical University of Denmark, DTU Informatics, Lyngby, Denmark, bw@imm.dtu.dk

Appendix G

Knowledge Discovery in Neuroinformatics

Bartłomiej Wilkowski. Knowledge Discovery in Neuroinformatics. *Presented at: Medical Informatics in a United and Healthy Europe*, 150:589, Sarajevo 2009. Published.

Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-589

Knowledge Discovery in Neuroinformatics

Bartlomiej WILKOWSKI [a,b,1]
[a] *DTU Informatics, Technical University of Denmark, Lyngby, Denmark*
[b] *Center for Integrated Molecular Brain Imaging, Copenhagen, Denmark*

Abstract. Traditionally, the process of turning data into biomedical knowledge has involved "manual" meta-analyses of results reported in journals. Since the amount of scientific data produced in neuroscience today increases dramatically, the resultant expansion of the medical databases has created a significant potential for the design of new data modeling and information retrieval tools and services that enable faster data processing, analysis and dissemination among a highly interdisciplinary community of researchers.

Keywords. data sharing, neuroinformatics, medical ontology

This PhD study aims at the discussion and design of necessary tools, methods and software, which can help in turning data into biomedical knowledge (meta-analysis of results reported in journals and information retrieval) [1], efficient medical data modeling and integration of various databases and repositories facilitating the data exchange between researchers from the whole neuroscience field.

The Center for Integrated Molecular Brain Imaging (CIMBI) with which this PhD study is associated, has established a large database of behavioral, genetic and imaging data. The key challenge of the research to be carried out during the PhD study is to develop the methods for integration of the CIMBI and related distributed databases including literature, biobanks and DTU's functional imaging database "Brede" in order to create an intelligent service for efficient information retrieval. Such a service is likely to become important not only for extracting information but also for an assistance in various aspects of research such as discovery of new facts, identification of previously undiscovered associations followed by proposal of new functions [2].

Medical ontologies (e.g., UMLS, NeuroLex) and formal and statistical methods (e.g., LSA, NMF) are considered to be a key for both database integration as well as the development of a process which is referred to as "knowledge discovery". Such a process can be understood as a pipeline of methods and techniques which include: text-mining of the scientific publications and further information retrieval (keyword extraction) [3], automatic interpretation of findings, discovery of new relationships and even design of new experiments.

[1] Wilkowski, B., Szewczyk, M.M., Rasmussen, P.M. et al. (2009) Coordinate-based meta-analytic search for the SPM neuroimaging pipeline – The BredeQuery plugin for SPM5. In *Proceedings of the International Conference on Health Informatics (HEALTHINF 2009)*, INSTICC Press, 11–17.
[2] Krallinger, M., Valencia, A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biology* 6(7):224.
[3] Wilkowski, B., Szewczyk, M.M., Hansen, L.K. (2009) Bridging the gap between coordinate- and keyword-based search of neuroscientific databases by UMLS-assisted semantic keyword extraction. Presented at the *15th Annual Meeting of the Organization for Human Brain Mapping.*

[1] Corresponding Author: Richard Petersens Plads, 2800 Kgs. Lyngby, Denmark; E-mail: bw@imm.dtu.dk.

Appendix H

Bridging the gap between coordinate- and keyword-based search of neuroscientific databases by UMLS-assisted semantic keyword extraction

---

Bartłomiej Wilkowski, Marcin Szewczyk, Lars Kai Hansen. Bridging the gap between coordinate- and keyword- based search of neuroscientific databases by UMLS-assisted semantic keyword extraction. *Presented at: Human Brain Mapping*, NeuroImage, 47(S165):(pp. 3), San Francisco 2009. Published.

Bridging the gap between coordinate- and keyword- based search of neuroscientific databases by UMLS-assisted semantic keyword extraction

Bartłomiej Wilkowski, Marcin Marek Szewczyk, Lars Kai Hansen
Center for Integrated Molecular Brain Imaging
Technical University of Denmark, DTU Informatics
bw@imm.dtu.dk, msz@imm.dtu.dk, lkh@imm.dtu.dk

Introduction

The rapid growth of the neuroimaging literature brings the demand for integration, organization and dissemination among a highly interdisciplinary community of researchers, see e.g. (Wager, Lindquist & Kaplan 2007). Since functional localization in brain is normally represented in form of stereotaxic coordinates, it can be used directly in the process of retrieving related literature in a given functional context by the measure of coordinate distance. Current neuroimaging databases which provide coordinate-based search capabilities (Brede Database, BrainMap) contain relatively small number of publications (Szewczyk 2008), therefore an interconnection with more comprehensive bibliographical databases can extend the results pool. Recently, the BredeQuery plug-in for SPM pipeline was presented as a tool which enables coordinate-based querying of the Brede Database directly from SPM (Wilkowski, Szewczyk, Rasmussen, Hansen & Nielsen 2009). As an extension of the current plugin's functionality, we propose methods for integration of the Brede Database with the almost complete medical publication database - PubMed (`http://pubmed.org`).

Methods

The first step towards the integration of PubMed with the BredeQuery plug-in is efficient keyword extraction from abstracts returned by the Brede Database (Nielsen 2003) after coordinate-based searching. The extracted keywords can be later used for modelling PubMed's query. Keywords are concatenated using OR, AND logical operators. We are developing Semantic KEyword Extraction Pipeline for Medical Documents (SKEEPMED) web service for mapping terms from abstracts to the UMLS Metathesaurus concepts using the MetaMap08 program (Aronson 2001). As we focus on the neuroscientific literature, we extract two types of keywords: brain parts (**brain_part**) and other significant domain terms (**term**). The final query is constructed with the following structure: *(**brain_part_1** OR **brain_part_2** OR ...) AND (**term_1** AND **term_2** AND ...).*

Results

We queried the Brede Database with a test coordinate in Talairach (Talairach & Tournoux 1988) space (-8,1,9), which relates to the thalamus brain region. The highest ranked experiment returned by the database belongs to article "Neuroanatomical Correlates of Happiness, Sadness, and Disgust" by Richard D. Lane et al. (1997), in the following referred to as the "source". "Source" contains description of 10 experiments with a total of 90 reported coordinates. The following keywords were extracted by SKEEPMED: **brain_part** keywords: *cerebral cortex, thalamus, insula, frontal lobe* ; **term** keywords: *disgust, sadness, happiness, emotion*. The PubMed query based on these keywords returned the "source" and 20 additional articles closely related to the topic of "source". We inspected only articles published later than "source" (16), 8 of which do not contain any experiment coordinates. To investigate the relevance of the remaining 8 articles, we compared spatial closeness of experiment coordinates from these 8 articles with the "source" experiments by querying Brede Database. Experiment coordinates from 6 of them are located in a close neighbourhood of "source's" experiments. Results are presented in Table 1.

#	1. PubMed Article	2. Year	3. Position in Brede Database search
1	Neural correlates of heart rate variability during emotion (Lane RD et al.)	2009	#1 (70%)
2	Beyond disgust: impaired recognition of negative emotions prior to diagnosis in Huntington's disease (Johnson SA et al.)	2007	no coordinates reported
3	Disgust and happiness recognition correlate with anteroventral insula and amygdala volume respectively in preclinical Huntington's disease (Kipps CM et al.)	2007	#3 (20%)
4	An event related functional magnetic resonance imaging study of facial emotion processing in Asperger syndrome (Deeley Q et al.)	2007	–
5	Neurophysiological correlates of induced discrete emotions in humans: an individually oriented analysis (Aftanas LI et al.)	2006	no coordinates reported
6	Neurophysiological correlates of induced discrete emotions in humans: an individual analysis (Aftanas LI et al.)	2004	no coordinates reported
7	Functional neuroanatomy of emotions: a meta-analysis (Murphy FC et al.)	2003	no coordinates reported
8	Common and distinct neural responses during direct and incidental processing of multiple facial emotions (Winston JS et al.)	2003	#9 (20%)
9	A preferential increase in the extrastriate response to signals of danger (Surguladze SA et al.)	2003	#1 (10%)
10	Impaired facial emotion recognition in early-onset right mesial temporal lobe epilepsy (Meletti S et al.)	2003	no coordinates reported
11	Age-related differences in brain activation during emotional face processing (Gunning-Dixon FM et al.)	2003	–
12	An fMRI study of facial emotion processing in patients with schizophrenia (Gur RE et al.)	2002	#2 (60%)
13	Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI (Phan KL et al.)	2002	no coordinates reported
14	Deficits in recognition of emotional facial expression are still present in alcoholics after mid- to long-term abstinence (Kornreich C et al.)	2001	no coordinates reported
15	Activation of anterior paralimbic structures during guilt-related script-driven imagery (Shin LM et al.)	2000	#7 (50%)
16	Perception of emotion in frontotemporal dementia and Alzheimer disease (Lavenu I et al.)	1999	no coordinates reported

Table 1: Results of spatial closeness comparison between experiments from PubMed retrieved articles and "source". Column 3. shows the position of the best-matched "source's" experiment in the results list returned by Brede Database when querying a test article experiment coordinates. In the parentheses the percentage of matched "source's" experiments found in top 20 Brede Database results is shown.

Conclusions

Current neuroimaging databases are limited and we have discussed a new way of enhancing their usability. We use the highly refined information in the Brede database to form an informed query into the literature at large. The case story showed the viability of the approach and gives us confidence that coordinate based search can be combined with language processing for a productivity enhancing tool for all neuroimagers. Current work concerns quantitative testing of the method.

References

Aronson, A. (2001), 'Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program', *JOURNAL OF BIOMEDICAL INFORMATICS* **35**, 17–21.

Nielsen, F. Å. (2003), The Brede database: a small database for functional neuroimaging, *in* 'NeuroImage', Vol. 19, Elsevier. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19-22, 2003, New York, NY.

Szewczyk, M. M. (2008), Databases for neuroscience, Master's thesis, Technical University of Denmark.

Talairach, J. & Tournoux, P. (1988), *Co-planar Stereotaxic Atlas of the Human Brain*, Thieme Medical Publisher Inc, New York.

Wager, T. D., Lindquist, M. & Kaplan, L. (2007), 'Meta-analysis of functional neuroimaging data: current and future directions', *Social Cognitive and Affective Neuroscience* **2**(2), 150–158.

Wilkowski, B., Szewczyk, M. M., Rasmussen, P. M., Hansen, L. K. & Nielsen, F. Å. (2009), Coordinate-based meta-analytic search for the SPM neuroimaging pipeline - The BredeQuery plugin for SPM5, *in* 'HEALTHINF 2009'.

Bibliography

C.B. Ahlers, M. Fiszman, D. Demner-Fushman, F. Lang, and T.C. Rindflesch. Extracting semantic predications from medline citations for pharmacogenomics. In *Pacific Symposium on Biocomputing 2007: Maui, Hawaii, 3-7 January 2007*, page 209. World Scientific Pub Co Inc, 2006.

C.B. Ahlers, D. Hristovski, H. Kilicoglu, and T.C. Rindflesch. Using the literature-based discovery paradigm to investigate drug mechanisms. In *American Medical Informatics Association Annual Symposium*, pages 6–15, 2007.

K. Anyanwu, A. Maduko, and A. Sheth. SemRank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th international conference on World Wide Web*, pages 117–127. ACM, 2005. ISBN 1595930469.

A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *American Medical Informatics Association Annual Symposium*, pages 17–21, 2001.

A.R. Aronson and F.M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229, 2010. ISSN 1527-974X.

G.A. Ascoli. The Ups and Downs of Neuroscience Shares. *Neuroinformatics*, 4: 213–216(4), 2006.

T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001. ISSN 0036-8733.

O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267, 2004. ISSN 0305-1048.

M. Brett. The MNI brain and the Talairach atlas. *MRC Cognition and Brain Sciences Unit*, 1999.

G. Burns, D. Feng, and E. Hovy. Intelligent approaches to mining the primary research literature: techniques, systems, and examples. *Computational Intelligence in Medical Informatics*, pages 17–50, 2008.

L. Chen and C. Friedman. Extracting phenotypic information from the literature via natural language processing. In *Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004]*, volume 107, page 758. OCSL Press, 2004.

K.H. Cheung, E. Lim, M. Samwald, H. Chen, L. Marenco, M.E. Holford, T.M. Morse, P. Mutalik, G.M. Shepherd, and P.L. Miller. Approaches to neuroscience data integration. *Briefings in bioinformatics*, 10(4):345, 2009.

T. Cohen, R. Schvaneveldt, and D. Widdows. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2):240–256, 2010a.

T. Cohen, R.W. Schvaneveldt, and T.C. Rindflesch. Predication-based semantic indexing: permutations as a means to encode predications in semantic space. In *American Medical Informatics Association Annual Symposium*, pages 114–118, 2009.

T. Cohen, G.K. Whitfield, R.W. Schvaneveldt, K. Mukund, and T. Rindflesch. EpiphaNet: An Interactive Tool to Support Biomedical Discoveries. *Journal of biomedical discovery and collaboration*, 5:21, 2010b.

R. Cole and P. Bruza. A bare bones approach to literature-based discovery: An analysis of the Raynaud's/fish-oil and migraine-magnesium discoveries in semantic space. In *Discovery Science*, pages 84–98. Springer, 2005.

C.J. Crasto, L.N. Marenco, M. Migliore, B. Mao, P.M. Nadkarni, P. Miller, and G.M. Shepherd. Text mining neuroscience journal articles to populate neuroscience databases. *Neuroinformatics*, 1(3):215–237, 2003. ISSN 1539-2791.

C.J. Crasto, P. Masiar, and P.L. Miller. NeuroExtract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation. *Journal of the American Medical Informatics Association*, 14(3):355, 2007. ISSN 1527-974X.

P. Dupont, J. Callut, G. Dooms, JN Monette, and Y. Deville. Relevant subgraph extraction from random walks in a graph. *Research Report RR*, 380167, 2006.

A.C. Evans, D.L. Collins, P. Neelin, D. MacDonald, M. Kamber, and T.S. Marrett. Three-dimensional correlative imaging: applications in human brain mapping. *Functional neuroimaging: technical foundations*, pages 145–162, 1994.

M. Fiszman, T.C. Rindflesch, and H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL workshop on computational lexical semantics*, pages 76–83. Association for Computational Linguistics, 2004.

M. Fiszman, T.C. Rindflesch, and H. Kilicoglu. Summarizing drug information in Medline citations. In *American Medical Informatics Association Annual Symposium*, pages 254–258, 2006.

L.C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979. ISSN 0378-8733.

L. French and P. Pavlidis. Informatics in neuroscience. *Briefings in bioinformatics*, 8(6):446, 2007.

K.J. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171, 1994. ISSN 1097-0193.

S.S. Fuller, D. Revere, P.F. Bugni, and G.M. Martin. A knowledgebase system to enhance scientific discovery: Telemakus. *Biomedical Digital Libraries*, 1 (1):2, 2004.

D. Gardner, H. Akil, G.A. Ascoli, D.M. Bowden, W. Bug, D.E. Donohue, D.H. Goldberg, B. Grafstein, J.S. Grethe, A. Gupta, et al. The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, 6(3):149–160, 2008. ISSN 1539-2791.

M.D. Gordon and R.K. Lindsay. Toward discovery support systems: A replication, re-examination, and extension of swanson's work on literature-based discovery of a connection between raynaud's and fish oil. *Journal of the American Society for Information Science*, 47(2):116–128, 1996.

D. Hristovski, C. Friedman, T.C. Rindflesch, and B. Peterlin. Exploiting semantic relations for literature-based discovery. In *American Medical Informatics Association Annual Symposium*, pages 349–353, 2006.

D. Hristovski, C. Friedman, TC Rindflesch, and B. Peterlin. Literature-based knowledge discovery using natural language processing. *Literature-based Discovery*, pages 133–152, 2008.

D. Hristovski, A. Kastrin, B. Peterlin, and T. Rindflesch. Combining Semantic Relations and DNA Microarray Data for Novel Hypotheses Generation. *Linking Literature, Information, and Knowledge for Biology*, pages 53–61, 2010.

D. Hristovski, B. Peterlin, J.A. Mitchell, and S.M. Humphrey. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2-4):289–298, 2005.

D. Hristovski, J. Stare, B. Peterlin, and S. Dzeroski. Supporting discovery in medicine by association rule mining in Medline and UMLS. *Studies in health technology and informatics*, pages 1344–1348, 2001.

M.Y. Hsiao, C.C. Chen, and J.H. Chen. BrainKnowledge: A Human Brain Function Mapping Knowledge-Base System. *Neuroinformatics*, pages 1–19, 2010. ISSN 1539-2791.

H. Kilicoglu, M. Fiszman, A. Rodriguez, D. Shin, AM Ripple, and TC Rindflesch. Semantic MEDLINE: a web application to manage the results of PubMed searches. In *Proceedings of the Third International Symposium for Semantic Mining in Biomedicine*, pages 69–76, 2008.

M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biology*, 6(7):224, 2005. ISSN 1465-6906.

A.R. Laird, J.L. Lancaster, and P.T. Fox. BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics*, 3:65–77, 2005.

C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 167–170, New York, NY, USA, 2006. ACM Press. ISBN 1595932496.

J.L. Lancaster, D. Tordesillas-Gutiérrez, M. Martinez, F. Salinas, A. Evans, K. Zilles, J.C. Mazziotta, and P.T. Fox. Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Human Brain Mapping*, January 2007. ISSN 1065-9471.

X.L. Li, Y.Q. Cai, H. Qin, and Y.J. Wu. Therapeutic effect and mechanism of proanthocyanidins from grape seeds in rats with TNBS-induced ulcerative colitis. *Canadian journal of physiology and pharmacology*, 86(12):841–849, 2008. ISSN 0008-4212.

J. Lin and W.J. Wilbur. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8(1):423, 2007. ISSN 1471-2105.

Y. Liu and G.A. Ascoli. Value Added by Data Sharing: Long-Term Potentiation of Neuroscience Research: A Commentary on the 2007 SfN Satellite Symposium on Data Sharing. *Neuroinformatics*, 5:143–145(3), 2007.

Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 64. NIH Public Access, 2006.

Y. Lussier and C. Friedman. BiomedLEE: a natural-language processor for extracting and representing phenotypes, underlying molecular mechanisms and their relationships. *ISMB: 2007*, 2007.

E.A. Maguire and C.J. Mummery. Differential modulation of a common memory retrieval network revealed by positron emission tomography. *Hippocampus*, 9 (1):54–61, 1999. ISSN 1098-1063.

H.M. Müller, A. Rangarajan, T.K. Teal, and P.W. Sternberg. Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics*, 6(3):195–204, 2008. ISSN 1539-2791.

F.Å. Nielsen. The Brede database: a small database for functional neuroimaging. In *NeuroImage*, volume 19. Elsevier, jun 2003. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19-22, 2003, New York, NY.

F.Å. Nielsen. Brede Wiki: Neuroscience data structured in a wiki. *Cerebral Cortex*, 15(2):166–169, 2004.

F.Å. Nielsen and L.K. Hansen. Finding related functional neuroimaging volumes. *Artificial Intelligence in Medicine*, 30(2):141–151, 2004.

S. Ogawa, D.W. Tank, R. Menon, J.M. Ellermann, S.G. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):5951, 1992.

T. Otsuka, T. Togo, N. Sugiyama, K. Uehara, A. Yoshimi, A. Karashima, H. Shioya, and Y. Hirayasu. Perospirone augmentation of paroxetine in treatment of refractory obsessive-compulsive disorder with depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(2):564–566, 2007.

T.C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6): 462–477, 2003. ISSN 1532-0464.

M. Samwald, H. Chen, A. Ruttenberg, E. Lim, L. Marenco, P. Miller, G. Shepherd, and K.H. Cheung. Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience. *Artificial intelligence in medicine*, 48(1): 21–28, 2010. ISSN 0933-3657.

L. Smith, T. Rindflesch, and W.J. Wilbur. MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, 2004a. ISSN 1367-4803.

S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219, 2004b. ISSN 1053-8119.

C. Sneiderman, D. Demner-Fushman, M. Fiszman, G. Rosemblat, F.M. Lang, D. Norwood, and T.C. Rindflesch. Semantic processing to enhance retrieval of diagnosis citations from Medline. In *American Medical Informatics Association Annual Symposium*, page 1104, 2006.

P. Srinivasan and B. Libbus. Mining medline for implicit links between dietary substances and diseases. *Bioinformatics*, 20(suppl 1):i290, 2004.

D.R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7, 1986a. ISSN 0031-5982.

D.R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2): 103–118, 1986b.

D.R. Swanson and N.R. Smalheiser. Undiscovered public knowledge: a ten-year update. In *KDD'96*, 1996.

D.R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2):183–203, 1997.

J Talairach and P. Tournoux. *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme Medical Publisher Inc, New York, January 1988. ISBN 0865772932.

J.L. Teeters, K.D. Harris, K.J. Millman, B.A. Olshausen, and F.T. Sommer. Data sharing for computational neuroscience. *Neuroinformatics*, 6(1):47–55, March 2008. ISSN 1559-0089.

D. Van Essen, J. Dickson, J. Harwell, and D.W.C. Hanlon. SumsDB: online access to surface-based representations of cerebral and cerebellar cortex in primates and rodents. In *Human Brain Project Annual Meeting, Bethesda, MD, USA*, 2004.

J.D. Van Horn and C.A. Ball. Domain-specific data sharing in neuroscience: what do we have to learn from each other? *Neuroinformatics*, 6(2):117–121, June 2008. ISSN 1559-0089.

T.D. Wager, M. Lindquist, and L. Kaplan. Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience*, 2(2):150–158, 2007.

T.D. Wager, M.A. Lindquist, T.E. Nichols, H. Kober, and J.X. Van Snellenberg. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage*, 45(1):S210–S221, 2009. ISSN 1053-8119.

M. Weeber, H. Klein, A.R. Aronson, J.G. Mork, LT De Jong-van Den Berg, and R. Vos. Text-based discovery in biomedicine: the architecture of the DAD–system. In *American Medical Informatics Association Annual Symposium*, pages 903–907, 2000.

M. Weeber, H. Klein, L. de Jong-van den Berg, and R. Vos. Using concepts in literature-based discovery: Simulating swanson's raynaud–fish oil and migraine–magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557, 2001.

H. Zhang, M. Fiszman, D. Shin, C.M. Miller, G. Rosemblat, and T.C. Rindflesch. Degree centrality for semantic abstraction summarization of therapeutic studies. *Journal of Biomedical Informatics*, 44(5):830 – 838, 2011. ISSN 1532-0464.